PRECONDITIONING OF DISCONTINUOUS GALERKIN METHODS FOR SECOND ORDER ELLIPTIC PROBLEMS

A Dissertation

by

VESELIN ASENOV DOBREV

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2007

Major Subject: Mathematics

PRECONDITIONING OF DISCONTINUOUS GALERKIN METHODS FOR SECOND ORDER ELLIPTIC PROBLEMS

A Dissertation

by

VESELIN ASENOV DOBREV

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Raytcho Lazarov
Committee Members,	Marvin Adams
	James Bramble
	Joseph Pasciak
Head of Department,	Al Boggess

December 2007

Major Subject: Mathematics

ABSTRACT

Preconditioning of Discontinuous Galerkin Methods
for Second Order Elliptic Problems. (December 2007)
Veselin Asenov Dobrev, B.S., Sofia University
Chair of Advisory Committee: Dr. Raytcho Lazarov

We consider algorithms for preconditioning of two discontinuous Galerkin (DG) methods for second order elliptic problems, namely the symmetric interior penalty (SIPG) method and the method of Baumann and Oden.

For the SIPG method we first consider two-level preconditioners using coarse spaces of either continuous piecewise polynomial functions or piecewise constant (discontinuous) functions. We show that both choices give rise to uniform, with respect to the mesh size, preconditioners. We also consider multilevel preconditioners based on the same two types of coarse spaces. In the case when continuous coarse spaces are used, we prove that a variable V-cycle multigrid algorithm is a uniform preconditioner. We present numerical experiments illustrating the behavior of the considered preconditioners when applied to various test problems in three spatial dimensions. The numerical results confirm our theoretical results and in the cases not covered by the theory show the efficiency of the proposed algorithms.

Another approach for preconditioning the SIPG method that we consider is an algebraic multigrid algorithm using coarsening based on element agglomeration which is suitable for unstructured meshes. We also consider an improved version of the algorithm using a smoothed aggregation technique. We present numerical experiments using the proposed algorithms which show their efficiency as uniform preconditioners.

For the method of Baumann and Oden we construct a preconditioner based on an orthogonal splitting of the discrete space into piecewise constant functions and

functions with zero average over each element. We show that the preconditioner is uniformly spectrally equivalent to an appropriate symmetrization of the discrete equations when quadratic or higher order finite elements are used. In the case of linear elements we give a characterization of the kernel of the discrete system and present numerical evidence that the method has optimal convergence rates in both L^2 and H^1 norms. We present numerical experiments which show that the convergence of the proposed preconditioning technique is independent of the mesh size.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Prof. Raytcho Lazarov, for his continued support and belief in me throughout my graduate studies. I am grateful to him for being my teacher and mentor, for the interesting and fruitful topic of my dissertation, and for all the shared ideas and given suggestions and directions during our numerous discussions. His constant encouragement and enthusiasm have been an inspiration to me.

I would like to thank Prof. James Bramble and Prof. Joseph Pasciak, not only for serving as members of my graduate committee, but also for being my teachers. Their lectures have introduced me to many new topics and have expanded greatly my knowledge in the area of numerical analysis and mathematics in general. I am grateful to Prof. Pasciak for thoroughly reading this dissertation.

I would like to thank Prof. Marvin Adams for serving as a member of my graduate committee and for introducing me to the interesting area of computational particle transport.

I would like to thank Prof. Panayot Vassilevski and Prof. Ludmil Zikatanov for giving me the opportunity to work with them, for sharing their ideas with me, and for their guidance and advice. I am especially grateful to Prof. Vassilevski for being my mentor during my summer internships at the Lawrence Livermore National Laboratory which have been a great learning experience for me.

I would like to thank all my teachers from Texas A&M University and Sofia University for sharing their knowledge with me.

I thank my parents and family for their encouragement and support throughout the years.

I thank all my friends for making my student years such an enjoyable time and

for all the interesting conversations and discussions on both mathematical and non-mathematical subjects.

Last but not least, I would like to thank the Department of Mathematics for giving me the opportunity to pursue my doctoral degree at Texas A&M University.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

TABLE Page

LIST OF FIGURES

CHAPTER I

INTRODUCTION

In recent years, the discontinuous Galerkin (DG) finite element methods have become a popular tool for the discretization of partial differential equations which can be applied to a large variety of problems. In contrast to the standard (Galerkin) finite element methods where the discrete spaces are chosen to preserve the natural continuity properties of the underlying PDE, the DG methods use discrete spaces with relaxed or no continuity restrictions across element boundaries. In order to impose the lost natural continuity weakly (without which one cannot expect good approximation properties) the DG methods introduce modifications to the bilinear and linear forms of the variational formulation of the problem.

Although the first DG methods were used to discretize hyperbolic equations, there are a number of DG methods for second-order elliptic equations. A unified analysis of a large class of such methods is presented in [2]. Among those are: the method of Babuška and Zlámal [3]; the symmetric interior penalty (IP or SIPG) method [18], [27], [1]; the method of Bassi and Rebay [4]; the method of Brezzi et al. [12]; the local discontinuous Galerkin (LDG) method [15]; the method of Baumann and Oden [5]; and the non-symmetric interior penalty (NIPG) method [22]. Table 1.1, taken from [2], summarizes some of the properties and error estimates for these DG methods. The columns "Symm." and "Cons." show if the bilinear form of the method is symmetric and if the method is consistent (i. e. the exact solution satisfies the variational equation of the method). The columns "H^1" and "L^2" give the order of the error estimate in H^1-like and L^2 norms when the discrete spaces use polynomials

---

This dissertation follows the style of the SIAM Journal on Numerical Analysis.

Table 1.1. Properties of DG methods for second-order elliptic problems

Method	Symm.	Cons.	H^1	L^2
Babuška–Zlámal [3]	✓	×	h^p	h^{p+1}
SIPG [18]	✓	✓	h^p	h^{p+1}
Bassi–Rebay [4]	✓	✓	$[h^p]$	$[h^{p+1}]$
Brezzi et al. [12]	✓	✓	h^p	h^{p+1}
LDG [15]	✓	✓	h^p	h^{p+1}
NIPG [22]	×	✓	h^p	h^p
Baumann–Oden [5], $p \geq 2$	×	✓	h^p	h^p

of degree p (the $[\cdot]$ brackets indicate that the estimates are in certain seminorms).

An important practical aspect of any discretization method is the ability to efficiently solve the resulting system of algebraic equations which can easily have a very large number of unknowns especially in two and three spatial dimensions. The use of DG discrete spaces further increases the number of unknowns compared to standard discretizations. This emphasizes even further the importance of efficient solvers for DG methods. A standard approach for solving large linear systems of equations with sparse matrices is to use an iterative method (e. g. PCG) coupled with a preconditioner to accelerate the convergence rate of the iteration. The multigrid methods are widely considered to be the best choice for the construction of preconditioners when a hierarchy of discretizations is available. It is well known that when applied to systems arising from standard finite element discretizations of second-order elliptic problems, the multigrid preconditioners are optimal and therefore they are a natural choice for DG methods as well.

The first work on multigrid preconditioning of DG discretizations that we are

aware of, is [19] where the authors introduce and analyze a variable V-cycle algorithm for the SIPG method. They show that the resulting preconditioner is spectrally equivalent to the matrix of the linear system. The proof is based on the abstract theory from [6] and requires a weak elliptic regularity assumption for the homogeneous continuous problem: $\|u\|_{1+\alpha} \leq C\|\Delta u\|_{-1+\alpha}$, for some $\alpha \in (\frac{1}{2}, 1]$.

Another recent work is [9] where V-, W-, and F-cycle multigrid algorithms for the SIPG method are considered. The algorithms are similar to the one in [19] in that they use the same sequence of coarse spaces. Assuming the same weak elliptic regularity, the authors prove that the energy norms of the error propagation operators are bounded by c/m^{α} when the number of smoothing steps m is sufficiently large: $m \geq m_0$. Their analysis is based on the theory developed by the authors in earlier papers and uses estimates in certain mesh-dependent scale of discrete norms.

Another recent work on preconditioning for the SIPG method is [13] where the authors consider multilevel Schwarz algorithms that give rise to uniform preconditioners. In contrast to [19], their analysis does not require regularity assumption. In addition the algorithm can be applied to problems on meshes with hanging nodes satistying mild grading conditions. The construction is based on a stable splitting of the discontinuous finite element space into conforming (continuous) subspace and a suitable non-conforming (discontinuous) subspace.

In Chapter III we consider algorithms for preconditioning of the SIPG method. First, we introduce two-level methods based on two choices for the coarse space both defined on the same triangulation as the DG discretization space $\mathcal{V}$: the first one consists of the continuous functions in $\mathcal{V}$, and the second one is the space of piecewise constant functions. We show that for both two-level preconditioners the energy norm of the error propagation operator is less than 1 uniformly in h. These results are contained in [16]. In the second part of Chapter III we introduce multigrid extensions

of the two-level preconditioners based on hierarchy of coarse spaces of the same type as the coarse spaces of the two-level methods. We prove that the variable V-cycle method using continuous coarse spaces gives rise to a spectrally equivalent preconditioner for the SIPG linear system which to the best of our knowledge is a new result. Our analysis is similar to the one used in [19]. We present numerical experiments in 3D that illustrate the theoretical results and investigate numerically some cases that are not covered by the analysis. The multilevel results are summarized in [17].

In Chapter IV we introduce algebraic multigrid (AMG) preconditioners for the SIPG method which can be used when the triangulation of the domain is unstructured. Our approach is to use AMGe based on element agglomeration introduced in [10], [20]. We use piecewise constant spaces on the coarse triangulations consisting of agglomerated elements and use the natural embeddings of these discontinuous spaces to define interpolation and coarse-level operators. We also consider a smoothed aggregation version using the ideas introduced in [25], [26]. The efficiency of the proposed AMG methods is investigated numerically and compared to the multigrid methods of Chapter III when possible.

In the last Chapter V we introduce and study a preconditioner for the non-symmetric method of Baumann and Oden. We start by studying the properties of the bilinear form of the method on the discrete DG space. Namely, we use an equivalent mixed formulation of the method to prove that the bilinear form satisfies an inf-sup condition in a suitable norm when quadratic or higher order elements are used. In the case of linear elements we characterize the kernel of the linear operator corresponding to the bilinear form, derive an equivalent form for the inf-sup condition and finally we show that it does not always hold with constant independent of h. We present numerical evidence that in this case (linear elements) the method has optimal convergence rates in both L^2 and H^1 norms. Also, we show that the norm used in the

inf-sup condition we proved defines a symmetric and positive definite preconditioner for a symmetrization of the non-symmetric system of the DG method. The action of the preconditioner requires the solution of a problem on the piecewise constant space with the SIPG bilinear form. If this smaller problem is solved exactly or replaced by an optimal preconditioner (e. g. using multigrid) the resulting preconditioner for the symmetrized system is also optimal. The results in this chapter are new and the preconditioning technique we introduced is the first optimal one that we know of. We conclude Chapter V with numerical experiments that confirm our analysis.

CHAPTER II

MODEL PROBLEM AND DG METHODS

In this chapter we introduce the model second order elliptic problem that we will consider. We also describe the general domain discretization and notation that we use. Then we proceed to introduce the discontinuous Galerkin methods that we will consider and establish some of their basic properties.

2.1. Model Second-Order Elliptic Problem

Let Ω be polyhedral domain in $\mathbb{R}^d$, $d = 2, 3$ with Lipschitz boundary and let $\mathbf{n}$ denote the outward unit normal vector to $\partial\Omega$. Assume that the boundary is decomposed in two disjoint components $\partial\Omega = \Gamma_D \cup \Gamma_N$, $\Gamma_D \cap \Gamma_N = \emptyset$ with Γ_D having positive boundary measure. We consider the following second-order elliptic boundary value problem:

$$\begin{aligned} -\nabla \cdot (a\nabla u) &= f \qquad \text{in } \Omega, \\ u &= u_D \quad \text{on } \Gamma_D, \\ (a\nabla u) \cdot \mathbf{n} &= u_N \quad \text{on } \Gamma_N. \end{aligned} \tag{2.1}$$

Here u is the unknown function and a, f, u_D, and u_N are given functions. We assume that $f \in L^2(\Omega)$, $u_D \in H^{1/2}(\Gamma_D)$, $u_N \in H^{-1/2}(\Gamma_N)$ and that the coefficient matrix $a \in (L^\infty(\Omega))^{d\times d}$ is symmetric, uniformly bounded and positive definite, that is there exist positive constants a_0 and a_1 such that

$$0 < a_0|\xi|^2 \le (a(x)\xi)\cdot\xi \le a_1|\xi|^2, \qquad \forall \xi \in \mathbb{R}^d, \text{ a. e. } x \in \Omega. \tag{2.2}$$

Let $\tilde{u}_D \in H^1(\Omega)$ be an extension of u_D inside the domain (i. e. $\tilde{u}_D|_{\Gamma_D} = u_D$) and define the space

$$H_0^1(\Omega;\Gamma_D) = \left\{ v \in H^1(\Omega) : u|_{\Gamma_D} = 0 \right\}.$$

It is well known that an equivalent variational formulation of the problem (2.1) is: find $u \in \tilde{u}_D + H_0^1(\Omega; \Gamma_D)$ such that

$$(a\nabla u, \nabla v) = (f, v) + \langle u_N, v \rangle_{\Gamma_N}, \qquad \forall v \in H_0^1(\Omega; \Gamma_D), \tag{2.3}$$

where $(\cdot, \cdot)$ denotes the inner product in $L^2(\Omega)$ and $L^2(\Omega)^d$, and $\langle \cdot, \cdot \rangle_{\Gamma_N}$ denotes the duality between $H^{-1/2}(\Gamma_N)$ and $H_{00}^{1/2}(\Gamma_N)$.

2.2. Domain Discretization

Let $\mathcal{T} = \{T\}$ be a simplicial discretization of the domain Ω into finite elements T that is the elements are triangles when $d = 2$ and tetrahedra when $d = 3$. We assume that $\mathcal{T}$ is a regular triangulation, that is the intersection of any two elements is either empty or a common vertex, edge, or face. We also assume that the elements are shape regular: there exists a constant γ such that

$$\frac{h_T}{\rho_T} \leq \gamma, \qquad \forall T \in \mathcal{T},$$

where h_T denotes the diameter of the element T and ρ_T — the diameter of the largest ball inscribed in T.

The set of all edges ($d = 2$) or faces ($d = 3$) of the elements in $\mathcal{T}$ will be denoted by $\mathcal{E}$ and we will refer to its elements as faces for both $d = 2$ and $d = 3$. Let $\mathcal{E}_i$ and $\mathcal{E}_b$ denote the sets of all interior and all boundary faces, respectively. We will assume that the Dirichlet boundary, Γ_D, is the union of a non-empty set of boundary faces which we will denote by $\mathcal{E}_D$. Similarly, $\mathcal{E}_N = \mathcal{E}_b \setminus \mathcal{E}_D$ will denote the faces where Neumann boundary condition is imposed. Thus, we have

$$\Gamma_D = \cup \mathcal{E}_D \qquad \Gamma_N = \cup \mathcal{E}_N.$$

For a given face $F \in \mathcal{E}$, $\mathcal{T}_F$ will denote the set of the elements that share the face F. Similarly, for an element $T \in \mathcal{T}$, $\mathcal{E}_T$ will denote the set of all faces of T. On the union of all faces in $\mathcal{E}$ we define the following piecewise constant function

$$h_{\mathcal{E}}|_F = |F|^{\frac{1}{d-1}}, \quad \forall F \in \mathcal{E},$$

where $|F|$ denotes the $\mathbb{R}^{d-1}$ dimensional measure of F. Since we assumed that the triangulation is simplicial and shape regular, if $F \in \mathcal{E}$ is one of the faces of an element $T \in \mathcal{T}$ then

$$c_0 h_T \leq h_{\mathcal{E}}|_F \leq c_1 h_T,$$

with constants c_0 and c_1 independent of T and F.

2.3. Function Spaces and Notation

We define the following "broken", with respect to the triangulation $\mathcal{T}$, Sobolev space:

$$H^s(\mathcal{T}) = \left\{ v \in L^2(\Omega) : v|_T \in H^s(T),\ \forall T \in \mathcal{T} \right\}, \qquad \text{for } s \geq 0.$$

We will consider the following discretization space of discontinuous piecewise polynomial functions of degree $r \geq 1$:

$$\mathcal{V} \equiv \mathcal{V}(\mathcal{T}, r) = \left\{ v \in L^2(\Omega) : v|_T \in P_r(T),\ \forall T \in \mathcal{T} \right\}.$$

Let $F \in \mathcal{E}_i$ be a face shared by two elements, T_i and T_j from $\mathcal{T}$ (see Figure 2.1); we denote the unit vector normal to F pointing from T_i to T_j by $\mathbf{n}$ (thus we choose and fix a direction for each interior face). For boundary face F, $\mathbf{n}$ will denote the unit normal vector pointing outside of Ω. Let w be a function defined on both sides of F as w_i and w_j from the side of the elements T_i and T_j, respectively. For example, w can be the trace of a function in $H^s(\mathcal{T})$, $s > 1/2$, or the normal derivative, $\nabla u \cdot \mathbf{n}$, of a

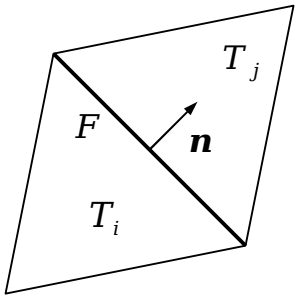

Fig. 2.1. Interior edge shared by two elements.

function $u \in H^s(\mathcal{T})$, $s > 3/2$. For such w we define the jump and average operators:

$$[\![w]\!] = w_i - w_j \qquad \text{and} \qquad \{w\} = \frac{1}{2}(w_i + w_j).$$

Thus, the direction for the jump $[\![w]\!]$ is determined by our choice of the direction of the normal vector $\mathbf{n}$. For a boundary face F and w defined only on the interior side of it as w_i we set

$$[\![w]\!] = w_i \qquad \text{and} \qquad \{w\} = w_i.$$

For a function $u \in H^1(\mathcal{T})$ we will use the notation ∇u to denote the element by element derivative of u instead of its distributional derivative:

$$\nabla u \in L^2(\Omega) \ : \ (\nabla u)|_T = \nabla(u|_T), \qquad \forall T \in \mathcal{T}, \quad \forall u \in H^1(\mathcal{T}).$$

For functions $p, q \in L^2(\cup S)$ defined on the union of some set of faces S, e.g. $S = \mathcal{E}$, we will use the following notation

$$\langle p, q \rangle_S = \sum_{F \in S} \int_F p\, q = \int_{\cup S} p\, q.$$

2.4. Discontinuous Galerkin Methods

In order to define the DG methods considered in this section we will assume that the coefficients a and u_N are smooth. Namely, we assume that $u_N \in L^2(\Gamma_N)$ and that a is piecewise $(W^1_\infty)^{d\times d}$ with respect to the triangulation $\mathcal{T}$. This allows us to define traces of $(a\nabla u)$ when $u \in H^s(\mathcal{T})$ for $s > 3/2$. Let $u, v \in H^s(\mathcal{T})$ for some $s > 3/2$ and define the DG bilinear form

$$\begin{aligned}\mathcal{A}(u,v) = (a\nabla u, \nabla v) &- \langle \{a\nabla u \cdot \mathbf{n}\}, [\![v]\!] \rangle_{\mathcal{E}_i \cup \mathcal{E}_D} \\ &+ \sigma \langle \{a\nabla v \cdot \mathbf{n}\}, [\![u]\!] \rangle_{\mathcal{E}_i \cup \mathcal{E}_D} + \left\langle \kappa h_{\mathcal{E}}^{-1} a_{\mathcal{E}} [\![u]\!], [\![v]\!] \right\rangle_{\mathcal{E}_i \cup \mathcal{E}_D}\end{aligned}$$

and the linear form

$$\mathcal{L}(v) = (f,v) + \langle u_N, v\rangle_{\mathcal{E}_N} + \sigma \langle u_D, a\nabla v \cdot \mathbf{n}\rangle_{\mathcal{E}_D} + \left\langle \kappa h_{\mathcal{E}}^{-1} a_{\mathcal{E}} u_D, v \right\rangle_{\mathcal{E}_D},$$

where the choices of $\sigma = \pm 1$ and $\kappa \geq 0$ will define different DG methods and $a_{\mathcal{E}}$ is a restriction of the coefficient a to the faces; one possible choice for $a_{\mathcal{E}}$ is

$$a_{\mathcal{E}} = \{a\mathbf{n} \cdot \mathbf{n}\}.$$

It is not hard to see that the definition is independent of our choices for the directions of the normal vectors $\mathbf{n}$ on the interior faces. The following choices for σ and κ define some of the well known DG methods:

- $\sigma = -1$ and $\kappa \geq \kappa_0$ sufficiently large define the symmetric interior penalty (IP or SIPG) method (see [27], [1], [2]).
- $\sigma = +1$ and $\kappa > 0$ define the non-symmetric interior penalty (NIPG) method (see [22], [2]).
- $\sigma = +1$ and $\kappa = 0$ define the method of Baumann and Oden (see [5], [21], [22], [2]).

Using the defined bilinear and linear forms the discrete DG problem can be written as: find $u \in \mathcal{V}$ such that

$$\mathcal{A}(u, v) = \mathcal{L}(v), \qquad \forall v \in \mathcal{V}.$$

Next, we look at some of the basic properties of these methods.

Proposition 1. *All three of the above methods are consistent. That is, if the exact solution U to (2.1) is in $H^s(\Omega)$ for some $s > 3/2$ then we have*

$$\mathcal{A}(U, v) = \mathcal{L}(v), \qquad \forall v \in H^s(\mathcal{T}).$$

Proof. Let $v \in H^s(\mathcal{T})$ be arbitrary test function. Multiplying the first equation in (2.1) by v, integrating over Ω and then integrating by parts over each element separately gives

$$(f, v) = (-\nabla \cdot (a\nabla U), v) = \sum_{T \in \mathcal{T}} (a\nabla U, \nabla v)_T - \langle a\nabla U \cdot \mathbf{n}_T, v \rangle_{\partial T}$$

where $\mathbf{n}_T$ is the unit outward vector normal to ∂T. Since $a\nabla U \in H(\mathrm{div}; \Omega)$ its normal component has no jump through interior faces and using the basic algebraic equality

$$(\mathbf{w}_1 \cdot \mathbf{n}_1) z_1 + (\mathbf{w}_2 \cdot \mathbf{n}_2) z_2 = \{\mathbf{w} \cdot \mathbf{n}\} [\![z]\!] + [\![\mathbf{w} \cdot \mathbf{n}]\!] \{z\}$$

we can rewrite the second term in the sum as

$$\sum_{T \in \mathcal{T}} \langle a\nabla U \cdot \mathbf{n}_T, v \rangle_{\partial T} = \langle \{a\nabla U \cdot \mathbf{n}\}, [\![v]\!] \rangle_{\mathcal{E}_i \cup \mathcal{E}_D} + \langle u_N, v \rangle_{\mathcal{E}_N}$$

where we also used the definition of jump and average on $\mathcal{E}_D$ and the last equation in (2.1) on $\mathcal{E}_N$. Thus, we arrive at the equality

$$(a\nabla U, \nabla v) - \langle \{a\nabla U \cdot \mathbf{n}\}, [\![v]\!] \rangle_{\mathcal{E}_i \cup \mathcal{E}_D} = (f, v) + \langle u_N, v \rangle_{\mathcal{E}_N}$$

which combined with the obvious equalities (since $[\![U]\!] = 0$ on $\mathcal{E}_i$, and $[\![U]\!] = U = u_D$ on $\mathcal{E}_D$)

$$\sigma \langle \{a\nabla v \cdot \mathbf{n}\}, [\![U]\!] \rangle_{\mathcal{E}_i \cup \mathcal{E}_D} = \sigma \langle u_D, a\nabla v \cdot \mathbf{n} \rangle_{\mathcal{E}_D}$$

$$\left\langle \kappa h_{\mathcal{E}}^{-1} a_{\mathcal{E}} [\![U]\!], [\![v]\!] \right\rangle_{\mathcal{E}_i \cup \mathcal{E}_D} = \left\langle \kappa h_{\mathcal{E}}^{-1} a_{\mathcal{E}} u_D, v \right\rangle_{\mathcal{E}_D},$$

gives the desired consistency. □

We now look at the boundedness and coercivity properties of the bilinear form $\mathcal{A}(\cdot,\cdot)$ with respect to the norm

$$|||v|||^2 = |||v|||_{a,\kappa}^2 = (a\nabla v, \nabla v) + \left\langle \kappa h_{\mathcal{E}}^{-1} a_{\mathcal{E}} [\![v]\!], [\![v]\!] \right\rangle_{\mathcal{E}_i \cup \mathcal{E}_D}, \qquad \forall v \in H^1(\mathcal{T}). \tag{2.4}$$

We begin with the following discrete estimate:

Lemma 1. *Let T be a non-degenerate simplex in $\mathbb{R}^d$ and F — one of its faces. If $\phi \in P_r(T)$ is a polynomial of degree r then*

$$\|\phi\|_{0,F}^2 \leq C \frac{|F|}{|T|} \|\phi\|_{0,T}^2$$

where C depends only on d and r. Here $\|\cdot\|_{0,S}$ and $|S|$ denote the norm in $L^2(S)$ and the measure of S, respectively.

Proof. Let $\hat{T}$ be a reference simplex in $\mathbb{R}^d$ and let $\hat{F}$ be one of its faces. Let

$$x = G(\hat{x}) = B\hat{x} + b$$

be the affine transformation that transforms $\hat{T}$ to T and $\hat{F}$ to F. Since G is affine, the function

$$\hat{\phi}(\hat{x}) = \phi(G(\hat{x}))$$

is a polynomial of degree r and therefore

$$\|\hat{\phi}\|_{0,\hat{F}}^2 \leq C \|\hat{\phi}\|_{0,\hat{T}}^2$$

because $\|\cdot\|_{0,\hat{F}}$ is a seminorm in the finite dimensional space $P_r(\hat{T})$. The Jacobians J_F and J_T of the transformations $\hat{F} \to F$ and $\hat{T} \to T$ are constant and we can express them from the equalities

$$|F| = \int_F 1 dS = \int_{\hat{F}} 1 |J_F| d\hat{S} = |J_F||\hat{F}|$$
$$|T| = \int_F 1 dx = \int_{\hat{T}} 1 |J_T| d\hat{x} = |J_T||\hat{T}|.$$

Thus, we have (even for $\phi|_F \in L^2(F)$, $\phi \in L^2(T)$)

$$\|\phi\|^2_{0,F} = |J_F|\|\hat{\phi}\|^2_{0,\hat{F}} = \frac{|F|}{|\hat{F}|}\|\hat{\phi}\|^2_{0,\hat{F}} \qquad \|\phi\|^2_{0,T} = |J_T|\|\hat{\phi}\|^2_{0,\hat{T}} = \frac{|T|}{|\hat{T}|}\|\hat{\phi}\|^2_{0,\hat{T}} \tag{2.5}$$

and therefore

$$\|\phi\|^2_{0,F} = \frac{|F|}{|\hat{F}|}\|\hat{\phi}\|^2_{0,\hat{F}} \le C\frac{|F|}{|\hat{F}|}\|\hat{\phi}\|^2_{0,\hat{T}} = C\frac{|F|}{|\hat{F}|}\frac{|\hat{T}|}{|T|}\|\phi\|^2_{0,T}.$$

□

The next lemma is well known in the case of a slowly varying coefficient a with small or no jumps and when the elements are shape regular. We derive a slightly more general estimate:

Lemma 2. *Let $F \in \mathcal{E}$ be one of the faces of the element $T \in \mathcal{T}$, $v \in L^2(F)$ and $u \in P_r(T)$. Assume that the matrix coefficient a satisfies*

$$\alpha_T A_T \xi \cdot \xi \le a(x)\xi \cdot \xi \le A_T \xi \cdot \xi, \qquad \forall \xi \in \mathbb{R}^d, \ \text{a. e. } x \in T,$$

where $\alpha_T > 0$ is a constant and A_T is a constant s. p. d. matrix. Then for any $\kappa > 0$ we have

$$\int_F (a\nabla u \cdot \mathbf{n})v \le \frac{C}{\kappa}\frac{h_{\mathcal{E}}|_F}{\alpha_T}\frac{|F|}{|T|}(a\nabla u, \nabla u)_T + \frac{\kappa}{4}\left\langle h_{\mathcal{E}}^{-1}(a\mathbf{n}\cdot\mathbf{n})v, v\right\rangle_F,$$

with a constant C that depends only on d and r.

Proof. Since the matrix coefficient a is s. p. d. we can define its square root $a^{\frac{1}{2}}$ and write

$$\int_F (a\nabla u \cdot \mathbf{n})v = \int_F (a^{\frac{1}{2}}\nabla u)\cdot(a^{\frac{1}{2}}\mathbf{n})v \le \int_F |a^{\frac{1}{2}}\nabla u||a^{\frac{1}{2}}\mathbf{n}|\,|v|$$
$$\le \|\beta^{-\frac{1}{2}}|a^{\frac{1}{2}}\nabla u|\|_{0,F}\|\beta^{\frac{1}{2}}|a^{\frac{1}{2}}\mathbf{n}|v\|_{0,F} \le \frac{1}{2\beta}\||a^{\frac{1}{2}}\nabla u|\|^2_{0,F} + \frac{\beta}{2}\||a^{\frac{1}{2}}\mathbf{n}|v\|^2_{0,F},$$

where we take $\beta = \kappa/(2h_{\mathcal{E}})$. The second term becomes

$$\frac{\beta}{2}\||a^{\frac{1}{2}}\mathbf{n}|v\|^2_{0,F} = \frac{\kappa}{4}\left\langle h_{\mathcal{E}}^{-1}(a\mathbf{n}\cdot\mathbf{n})v, v\right\rangle_F$$

and the first term can be estimated

$$\frac{1}{2\beta}\||a^{\frac{1}{2}}\nabla u|\|^2_{0,F} = \frac{h_{\mathcal{E}}}{\kappa}\int_F a\nabla u\cdot\nabla u \le \frac{h_{\mathcal{E}}}{\kappa}\int_F A_T\nabla u\cdot\nabla u = \frac{h_{\mathcal{E}}}{\kappa}\sum_{i=1}^{d}\|(A_T^{\frac{1}{2}}\nabla u)_i\|^2_{0,F}.$$

Each component of $A_T^{\frac{1}{2}}\nabla u$ is a polynomial of degree $(r-1)$ and therefore we can apply Lemma 1

$$\frac{h_{\mathcal{E}}}{\kappa}\sum_{i=1}^{d}\|(A_T^{\frac{1}{2}}\nabla u)_i\|^2_{0,F} \le C\frac{h_{\mathcal{E}}}{\kappa}\frac{|F|}{|T|}\sum_{i=1}^{d}\|(A_T^{\frac{1}{2}}\nabla u)_i\|^2_{0,T} = C\frac{h_{\mathcal{E}}}{\kappa}\frac{|F|}{|T|}\int_T A_T\nabla u\cdot\nabla u$$
$$\le \frac{C}{\kappa}\frac{h_{\mathcal{E}}}{\alpha_T}\frac{|F|}{|T|}\int_T a\nabla u\cdot\nabla u.$$

to complete the proof. □

Corollary 1. *Let $\kappa > 0$ be arbitrary constant and let $a_{\mathcal{E}}$ be defined by*

$$a_{\mathcal{E}} = \{a\mathbf{n}\cdot\mathbf{n}\}.$$

Then the following estimate holds: $\forall u \in \mathcal{V}, \forall v \in H^s(\Omega), s > 3/2$

$$\langle\{a\nabla u\cdot\mathbf{n}\}, [\![v]\!]\rangle_{\mathcal{E}_i\cup\mathcal{E}_D} \le CC_1\frac{1}{\kappa}(a\nabla u,\nabla u) + \frac{\kappa}{4}\left\langle h_{\mathcal{E}}^{-1}a_{\mathcal{E}}[\![v]\!],[\![v]\!]\right\rangle_{\mathcal{E}_i\cup\mathcal{E}_D}$$

where

$$C_1 = \max_{T\in\mathcal{T},F\in\mathcal{E}_T}\left\{\frac{(h_{\mathcal{E}}|_F)|F|}{\alpha_T|T|}\right\}$$

and C depends only on d and r.

Remark 1. *It is clear that if the elements are shape regular as we assumed earlier then C_1 is independent of h_T since*

$$\frac{(h_{\mathcal{E}}|_F)|F|}{|T|} \simeq \frac{h_T h_T^{d-1}}{h_T^d} = 1.$$

Also, note that C_1 is independent of any jumps that the coefficient a may have across interior faces. For example, if it is piecewise constant w. r. t. the mesh $\mathcal{T}$ then $\alpha_T = 1$.

With the help of Corollary 1 we can derive the following

Proposition 2. *The bilinear form $\mathcal{A}(\cdot,\cdot)$ of all three methods is bounded on the discrete space $\mathcal{V}$ in the $|||\cdot|||_{a,\kappa}$ norm (with $\kappa = 1$ for the method of Baumann and Oden). The bilinear form of the SIPG method is coercive provided that $\kappa \geq \kappa_0 > 0$ is sufficiently large; the NIPG bilinear form is coercive for any $\kappa > 0$ (but the constant in the boundedness grows like $1/\kappa$ for small κ). All constants in these bounds are independent of h_T and any jumps of the coefficient a across interior faces.*

Proof. The boundedness follows easily from Corollary 1 and the coercivity of the NIPG form is obvious. In the case of the SIPG method we can estimate

$$\begin{aligned}
\mathcal{A}(v,v) &= (a\nabla v, \nabla v) + \left\langle \kappa h_{\mathcal{E}}^{-1} a_{\mathcal{E}} [\![v]\!], [\![v]\!] \right\rangle_{\mathcal{E}_i \cup \mathcal{E}_D} - 2 \left\langle \{a\nabla v \cdot \mathbf{n}\}, [\![v]\!] \right\rangle_{\mathcal{E}_i \cup \mathcal{E}_D} \\
&\geq (a\nabla v, \nabla v) + \left\langle \kappa h_{\mathcal{E}}^{-1} a_{\mathcal{E}} [\![v]\!], [\![v]\!] \right\rangle_{\mathcal{E}_i \cup \mathcal{E}_D} \\
&\quad - 2\left(CC_1 \frac{1}{\kappa} (a\nabla v, \nabla v) + \frac{\kappa}{4} \left\langle h_{\mathcal{E}}^{-1} a_{\mathcal{E}} [\![v]\!], [\![v]\!] \right\rangle_{\mathcal{E}_i \cup \mathcal{E}_D} \right) \\
&= (1 - 2CC_1/\kappa)(a\nabla v, \nabla v) + \frac{1}{2} \left\langle \kappa h_{\mathcal{E}}^{-1} a_{\mathcal{E}} [\![v]\!], [\![v]\!] \right\rangle_{\mathcal{E}_i \cup \mathcal{E}_D}.
\end{aligned}$$

Thus, if $2CC_1/\kappa < 1$ the coercivity will follow. □

Note that all the estimates in the Proposition are valid only on the *discrete space* $\mathcal{V}$.

CHAPTER III

PRECONDITIONING THE SIPG METHOD

In this chapter we consider preconditioning techniques for the symmetric interior penalty (SIPG) method. First, we consider two-level methods based on coarse spaces defined on the same mesh as the SIPG method. We prove that the resulting two-level preconditioners are uniformly (with respect to the mesh size) spectrally equivalent to the matrix of the discrete linear system and present numerical experiments that illustrate that. We then proceed to define natural multilevel extensions of the two-level methods. We analyze one of the multilevel methods using the abstract theory from [8]. We show that a variable V-cycle method gives rise to uniform preconditioner. We conclude the chapter with numerical experiments that test the presented multilevel methods.

3.1. Two-Level Methods

3.1.1. Description and Abstract Estimate

Consider a general subspace $\mathcal{V}_0 \subset \mathcal{V}$ of the discrete space where we will seek a solution and define the operators $A : \mathcal{V} \to \mathcal{V}$, $A_0 : \mathcal{V}_0 \to \mathcal{V}_0$, the L^2-orthogonal projector $Q : \mathcal{V} \to \mathcal{V}_0$, and the $\mathcal{A}(\cdot,\cdot)$ orthogonal projector $P : \mathcal{V} \to \mathcal{V}_0$ by:

$$
\begin{aligned}
(Au, v) &= \mathcal{A}(u, v), \quad \forall u, v \in \mathcal{V} \\
(A_0 u, v) &= \mathcal{A}(u, v), \quad \forall u, v \in \mathcal{V}_0 \\
(Qu, v) &= (u, v), \qquad \forall u \in \mathcal{V}, \quad \forall v \in \mathcal{V}_0 \\
\mathcal{A}(Pu, v) &= \mathcal{A}(u, v), \quad \forall u \in \mathcal{V}, \quad \forall v \in \mathcal{V}_0,
\end{aligned}
$$

where $(\cdot,\cdot)$ denotes the $L^2(\Omega)$ inner product. Also let $R : \mathcal{V} \to \mathcal{V}$ be a general smoother. For an operator $M : X \subset \mathcal{V} \to \mathcal{V}$ we will use $M^t : \mathcal{V} \to X$ to denote its transpose with respect to the $(\cdot,\cdot)$ inner product:

$$(M^t u, v) = (u, Mv), \qquad \forall u \in \mathcal{V}, \quad \forall v \in X.$$

Similarly, $M^* : \mathcal{V} \to X$ will denote the transpose of M with respect to the $\mathcal{A}(\cdot,\cdot)$ inner product:

$$\mathcal{A}(M^* u, v) = \mathcal{A}(u, Mv), \qquad \forall u \in \mathcal{V}, \quad \forall v \in X.$$

We consider the following two-level preconditioner $B : \mathcal{V} \to \mathcal{V}$ defined by the algorithm: given $g \in \mathcal{V}$ compute Bg by

1. pre-smoothing: $x = R^t g$
2. correction: $y = x + q$, where $q \in \mathcal{V}_0$ is the solution of
$$A_0 q = Q(g - Ax).$$
3. post-smoothing: $z = y + R(g - Ay)$.
4. then $Bg = z$.

In order to study the convergence properties of the two-level preconditioner B we introduce the so called error propagation operator $E = I - BA$ which can be written in the following product form

$$E \equiv I - BA = (I - RA)(I - P)(I - R^t A).$$

To see this, let $e \in \mathcal{V}$ be arbitrary and set $g = Ae$. Then using step 3 in the algorithm we get

$$Ee = e - Bg = e - z = e - y - R(g - Ay) = (e - y) - RA(e - y) = (I - RA)(e - y).$$

Now using step 2 we can write

$$e - y = e - x - q = e - x - A_0^{-1}Q(g - Ax) = (e - x) - A_0^{-1}QA(e - x) = (I - P)(e - x),$$

where we used the equality $A_0^{-1}QA = P$ which can be derived from the definitions. To this end, take arbitrary $u \in \mathcal{V}$ and $v \in \mathcal{V}_0$ and write

$$(QAu, v) = (Au, v) = \mathcal{A}(u, v) = \mathcal{A}(Pu, v) = (A_0Pu, v).$$

Since QAu and A_0Pu are in $\mathcal{V}_0$ and $v \in \mathcal{V}_0$ was arbitrary we conclude that $QAu = A_0Pu$ and therefore $QA = A_0P$ since $u \in \mathcal{V}$ was also arbitrary. Finally, using step 1 we find that

$$e - x = e - R^t g = e - R^t Ae = (I - R^t A)e$$

and consequently combining the above equalities we obtain the product form of E:

$$Ee = (I - RA)(e - y) = (I - RA)(I - P)(e - x) = (I - RA)(I - P)(I - R^t A)e.$$

A standard way of estimating the convergence properties of the linear iterative process

$$x^{i+1} = x^i + B(b - Ax^i)$$

for solving the equation $Ax = b$ given an arbitrary initial guess x^0, is to estimate the energy operator norm of E

$$\|E\|_A = \sup_v \frac{\|Ev\|_A}{\|v\|_A}, \qquad \text{where } \|v\|_A = (Av, v)^{1/2}.$$

Since we have that

$$x - x^{i+1} = x - x^i - B(Ax - Ax^i) = (I - BA)(x - x^i) = E(x - x^i),$$

proving a bound of the form

$$\|E\|_A < 1 \tag{3.1}$$

will guarantee the convergence of the linear iterative method. In addition such an estimate, combined with the fact (to be established in the next theorem) that E is symmetric and positive semi-definite with respect to the $\mathcal{A}(\cdot,\cdot)$ inner product will give the inequalities (with $\delta = \|E\|_A$)

$$0 \leq (AEv, v) \leq \delta(Av, v)$$

which are equivalent to

$$(1 - \delta)(Av, v) \leq (ABAv, v) \leq (Av, v),$$

that is the condition number of BA is bounded by $1/(1 - \delta)$. Thus, B is a good preconditioner for A that can be used in the preconditioned conjugate gradient (PCG) method to solve iteratively the equation $Ax = b$.

The next theorem gives sufficient conditions in the abstract setting presented so far that give rise to an estimate of the form (3.1).

Theorem 1. *The error propagation operator $E \equiv I - BA$ is symmetric and positive semi-definite in the inner product $\mathcal{A}(\cdot,\cdot)$. Also, if we assume that the following two conditions hold:*

1. *smoothing property: there exists $\omega > 0$ such that*

$$\frac{\omega}{\lambda}(v, v) \leq (\bar{R}v, v), \qquad \forall v \in \mathcal{V}, \tag{3.2}$$

where $\bar{R} = R + R^t - R^t AR$ *and* λ *is the largest eigenvalue of* A.

2. *approximation property: there exists an operator* $\widetilde{Q} : \mathcal{V} \to \mathcal{V}_0$ *such that*

$$\|v - \widetilde{Q}v\|^2 \leq \widehat{C}\lambda^{-1}(Av, v) \qquad \forall v \in \mathcal{V}. \tag{3.3}$$

then we have that

$$\|E\|_A \leq 1 - \frac{\omega}{\widehat{C}} < 1.$$

Proof. Let $K = I - RA$ then $\forall u, v \in \mathcal{V}$

$$\mathcal{A}(K^* u, v) = \mathcal{A}(u, Kv) = (u, (A - ARA)v) = ((A - AR^t A)u, v) = \mathcal{A}((I - R^t A)u, v)$$

or $K^* = I - R^t A$ and therefore the product form of E can be written as

$$E = K(I - P)K^*.$$

Since P is the $\mathcal{A}$-orthogonal projector onto $\mathcal{V}_0$ we have that $P^* = P$ and $P^2 = P$. The former of these two equalities easily implies the symmetry of E in the $\mathcal{A}$ inner product

$$E^* = (K^*)^*(I - P^*)K^* = K(I - P)K^* = E.$$

To see that E is positive semi-definite we note that $I - P = (I - P)^2$ and therefore

$$\mathcal{A}(Ev, v) = \mathcal{A}(K(I - P)^2 K^* v, v) = \mathcal{A}((I - P)K^* v, (I - P)K^* v) \geq 0.$$

The symmetry of E allows us to write its norm as

$$\|E\|_A = \sup_{v \in \mathcal{V}} \frac{\mathcal{A}(Ev, v)}{\mathcal{A}(v, v)} = \sup_{v \in \mathcal{V}} \frac{\mathcal{A}((I - P)K^* v, (I - P)K^* v)}{\mathcal{A}(v, v)} = \|(I - P)K^*\|_A^2.$$

Using the fact that $\|M\|_A = \|M^*\|_A$, we have

$$\|E\|_A = \|(I - P)K^*\|_A^2 = \|((I - P)K^*)^*\|_A^2 = \|K(I - P)\|_A^2.$$

Letting $v = Az$ in the smoothing property and using the equality

$$(AKz, Kz) = (Az, z) - (\bar{R}Az, Az)$$

we obtain the following equivalent form

$$(AKz, Kz) \leq (Az, z) - \frac{\omega}{\lambda}(Az, Az), \qquad \forall z \in \mathcal{V}.$$

We want to estimate $\|K(I-P)v\|_A^2$ for any $v \in \mathcal{V}$. Letting $z = (I-P)v$ in the above form of the smoothing property we get

$$\|K(I-P)v\|_A^2 = \|Kz\|_A^2 \leq (Az, z) - \frac{\omega}{\lambda}(Az, Az). \tag{3.4}$$

We will now use the approximation property to estimate (Az, z). Since $\widetilde{Q}z \in \mathcal{V}_0$ we have

$$\mathcal{A}(Pv, \widetilde{Q}z) = \mathcal{A}(v, \widetilde{Q}z), \qquad \text{or} \qquad \mathcal{A}(z, \widetilde{Q}z) = 0$$

and therefore

$$(Az, z) = (Az, z - \widetilde{Q}z) \leq (Az, Az)^{\frac{1}{2}} \|z - \widetilde{Q}z\| \leq (Az, Az)^{\frac{1}{2}} \widehat{C}^{\frac{1}{2}} \lambda^{-\frac{1}{2}} (Az, z)^{\frac{1}{2}}$$

or

$$(Az, z) \leq \widehat{C}\lambda^{-1}(Az, Az).$$

This is equivalent to

$$-\frac{\omega}{\lambda}(Az, Az) \leq \frac{\omega}{\widehat{C}}(Az, z)$$

which we can use in (3.4) to get

$$\|K(I-P)v\|_A^2 \leq \left(1 - \frac{\omega}{\widehat{C}}\right)(Az, z).$$

Notice that $(I - P)$ is the $\mathcal{A}$-orthogonal projection onto its image and therefore

$$(Az, z) = \|(I - P)v\|_A^2 \leq \|v\|_A^2$$

which combined with the estimate above gives

$$\|K(I - P)v\|_A^2 \leq \left(1 - \frac{\omega}{\widehat{C}}\right) \|v\|_A^2$$

or we get the final result of the theorem

$$\|E\|_A = \|K(I - P)\|_A^2 \leq 1 - \frac{\omega}{\widehat{C}}.$$

□

Note that the smoothing property in the theorem is the same as the assumption SM.1 used in [8].

3.1.2. Coarse Spaces and Analysis

We consider the following two coarse spaces as $\mathcal{V}_0$:

1. the space of continuous piecewise polynomials of the same degree r as in the space $\mathcal{V}$

   $$\mathcal{V}^c = \left\{v \in \mathcal{C}(\overline{\Omega}) : v|_T \in P_r(T),\ \forall T \in \mathcal{T}\right\} = \mathcal{C}(\overline{\Omega}) \cap \mathcal{V}.$$

   When restricted to this space the SIPG bilinear form simplifies to

   $$\begin{aligned}\mathcal{A}(u, v) = (a\nabla u, \nabla v) - \langle a\nabla u \cdot \mathbf{n}, v\rangle_{\mathcal{E}_D} \\ + \langle a\nabla v \cdot \mathbf{n}, u\rangle_{\mathcal{E}_D} + \left\langle \kappa h_{\mathcal{E}}^{-1} a_{\mathcal{E}} u, v\right\rangle_{\mathcal{E}_D}, \qquad \forall u, v \in \mathcal{V}^c.\end{aligned}$$

2. the space of piecewise constant functions

   $$\overline{\mathcal{V}} = \left\{v \in L^2(\Omega) : v|_T = \text{const},\ \forall T \in \mathcal{T}\right\}.$$

In this case the form simplifies to

$$\mathcal{A}(u,v) = \left\langle \kappa h_{\mathcal{E}}^{-1} a_{\mathcal{E}} \left[\!\left[u\right]\!\right], \left[\!\left[v\right]\!\right] \right\rangle_{\mathcal{E}_i \cup \mathcal{E}_D}, \qquad \forall u, v \in \overline{\mathcal{V}}.$$

In the analysis of the two two-level methods resulting from the two choices of the coarse space we will assume that the triangulation $\mathcal{T}$ is *globally* quasi-uniform, i. e. there exists a constant $c > 0$ such that

$$ch \le h_T \le h, \qquad \forall T \in \mathcal{T}, \qquad \text{where} \quad h = \max_{T \in \mathcal{T}} h_T.$$

We begin with the following

Proposition 3. *Let λ be the largest eigenvalue of the operator A. Then*

$$\lambda \equiv \sup_{v \in \mathcal{V}} \frac{(Av, v)}{(v, v)} \simeq h^{-2}.$$

Proof. Let $v \in \mathcal{V}$ be arbitrary then using Proposition 2 we get

$$\begin{aligned} \mathcal{A}(v,v) \le C \left|\!\left|\!\left| v \right|\!\right|\!\right|^2 &= (a \nabla v, \nabla v) + \left\langle \kappa h_{\mathcal{E}}^{-1} a_{\mathcal{E}} \left[\!\left[v\right]\!\right], \left[\!\left[v\right]\!\right] \right\rangle_{\mathcal{E}_i \cup \mathcal{E}_D} \\ &\le \sum_{T \in \mathcal{T}} a_1 |v|_{1,T}^2 + C\kappa h^{-1} \sum_{F \in \mathcal{E}_i \cup \mathcal{E}_D} a_1 \| \left[\!\left[v\right]\!\right] \|_{0,F}^2, \end{aligned}$$

where a_1 is the upper bound on a from (2.2). The first sum is easily estimated using an inverse inequality (cf. Theorem 3.2.6 in [14])

$$\sum_{T \in \mathcal{T}} a_1 |v|_{1,T}^2 \le C a_1 h^{-2} (v, v).$$

To estimate the second term we use Lemma 1

$$\| \left[\!\left[v\right]\!\right] \|_{0,F}^2 \le C \sum_{T \in \mathcal{T}_F} \| v|_T \|_{0,F}^2 \le C h^{-1} \sum_{T \in \mathcal{T}_F} \| v \|_{0,T}^2$$

and therefore after summation over the faces we get

$$\mathcal{A}(v,v) \le C h^{-2} (v, v)$$

which means that $\lambda \leq Ch^{-2}$. To see that the estimate is asymptotically sharp we consider the following $v \in \mathcal{V}$: let $T_0 \in \mathcal{T}$ be some fixed element and let $v|_{T_0} \in P_r(T_0)$ be arbitrary non-zero polynomial and let $v|_T \equiv 0$ for all other elements $T \in \mathcal{T} \setminus \{T_0\}$. For such v we have

$$\mathcal{A}(v,v) = (a\nabla v, \nabla v)_{T_0} + \sum_{F \in \mathcal{E}_{T_0} \setminus \mathcal{E}_N} \left\langle \kappa h_{\mathcal{E}}^{-1} a_{\mathcal{E}} v, v \right\rangle_F \simeq |v|^2_{1,T_0} + \kappa h^{-1} \sum_{F \in \mathcal{E}_{T_0} \setminus \mathcal{E}_N} \|v\|^2_{0,F}$$

From Theorems 3.1.2 and 3.1.3 in Ciarlet's book [14] one can derive the equivalence

$$|v|^2_{1,T_0} \simeq h^{d-2} |\hat{v}|^2_{1,\hat{T}}$$

where $\hat{T}$ is a reference simplex and $\hat{v}(\hat{x}) = v(G(\hat{x}))$ with G being the affine transformation from $\hat{T}$ to T_0. Also, as we saw in Lemma 1 we have

$$\|v\|^2_{0,F} \simeq h^{d-1} \|\hat{v}\|^2_{0,\hat{F}}$$

where $\hat{F} = G^{-1}(F)$. Thus we arrive at the norm equivalence

$$\mathcal{A}(v,v) \simeq h^{d-2} \left(|\hat{v}|^2_{0,\hat{T}} + \sum_{F \in \mathcal{E}_{T_0} \setminus \mathcal{E}_N} \|\hat{v}\|^2_{0,\hat{F}} \right).$$

On the other hand the L^2 norm of this special v is equivalent to (cf. Lemma 1)

$$(v,v) = \|v\|^2_{0,T_0} \simeq h^d \|\hat{v}\|^2_{0,\hat{T}}.$$

Now, it is easy to see that if we fix one such v we get

$$\lambda = \sup_{z \in \mathcal{V}} \frac{(Az,z)}{(z,z)} \geq \frac{(Av,v)}{(v,v)} \simeq h^{-2}.$$

□

In the next theorem we prove that the smoothing property (3.2) (which is independent of the coarse space $\mathcal{V}_0$) from Theorem 1 holds for some well known smoothers.

We use the abstract theory from [7], [8].

Theorem 2. *The smoothing property* (3.2) *holds with constant* ω *independent of* h *for any of the following (point) smoothers: scaled Jacobi, Gauss-Seidel, and symmetric Gauss-Seidel when applied to the SIPG bilinear form.*

Proof. Let $\{\phi_i\}_{i=1}^n$, $n = \dim \mathcal{V}$ be an ordering of the standard nodal basis in $\mathcal{V}$ and let $\mathcal{V}^i$ be the one-dimensional space spanned by ϕ_i. Let $P_i : \mathcal{V} \to \mathcal{V}^i$ be the $\mathcal{A}$ orthogonal projector onto $\mathcal{V}^i$

$$\mathcal{A}(P_i u, v) = \mathcal{A}(u, v), \qquad \forall u \in \mathcal{V},\ \forall v \in \mathcal{V}^i$$

and define the matrix σ with entries

$$\sigma_{ij} = \begin{cases} 0 & \text{if } P_i P_j = 0 \\ 1 & \text{otherwise.} \end{cases}$$

According to the abstract theory in [8], Section 8 we need to check the following two conditions:

1. there exists a constant $C_1 > 0$, independent of h such that

$$\|\sigma\|_\infty \equiv \max_{i=1\ldots n} \sum_{j=1}^{n} |\sigma_{ij}| \le C_1.$$

2. there exists a constant $C_2 > 0$, independent of h such that for each $v \in \mathcal{V}$ there is a decomposition $v = \sum_i v^i$, with $v^i \in \mathcal{V}^i$, satisfying

$$\sum_{i=1}^{n} \|v^i\|^2 \le C_2 \|v\|^2.$$

The condition $P_i P_j = 0$ is equivalent to the following conditions

$$\mathcal{A}(P_i P_j u, v) = 0, \quad \forall u, v \in \mathcal{V} \iff \mathcal{A}(P_j u, P_i v) = 0, \quad \forall u, v \in \mathcal{V}$$

the latter of which is clearly equivalent to

$$\mathcal{A}(u,v)=0, \quad \forall u\in\mathcal{V}^j, \forall v\in\mathcal{V}^i$$

that is the spaces $\mathcal{V}^i$ and $\mathcal{V}^j$ are $\mathcal{A}$-orthogonal which in this case is equivalent to $\mathcal{A}(\phi_i,\phi_j)=0$. Thus, σ has the sparsity pattern of the stiffness matrix. Since $\mathcal{A}(\phi_i,\phi_j)\neq 0$ is only possible when ϕ_i and ϕ_j correspond to (have support in) the same element or two elements with a common face, it is clear that we can choose C_1 depending only on the polynomial degree r and the dimension d. Since $\mathcal{V}$ is the direct sum of all spaces $\mathcal{V}^i$ there is only one possible decomposition for the second condition. Let $v=\sum_i v^i$, $v^i\in\mathcal{V}^i$ be that decomposition and let I_T denote the set of indices i for which ϕ_i has support in $T\in\mathcal{T}$. Note that each basis function has support in exactly one element so $\{I_T\}$ are a decomposition of the set $\{1,\dots,n\}$. Therefore

$$(v,v)=\sum_{T\in\mathcal{T}}(v,v)_T=\sum_{T\in\mathcal{T}}(v_T,v_T)_T$$

where

$$v_T(x)=\sum_{i\in I_T}v^i(x)=\begin{cases}v(x) & x\in T\\ 0 & x\notin T.\end{cases}$$

Let $\{\psi_i\}_{i=1}^m$ be the nodal basis on a reference simplex $\hat{T}$ and consider $\hat{v}=\sum_{i=1}^m\xi_i\psi_i$, $\xi_i\in\mathbb{R}$. Then if we denote $\hat{v}^i=\xi_i\psi_i$ we have

$$\sum_{i=1}^m\|\hat{v}^i\|_{\hat{T}}^2=\sum_{i=1}^m\|\xi_i\psi_i\|_{\hat{T}}^2=\sum_{i=1}^m\xi_i^2\|\psi_i\|_{\hat{T}}^2\simeq\sum_{i,j=1}^m\xi_i\xi_j(\psi_i,\psi_j)_{\hat{T}}=\|\hat{v}\|_{\hat{T}}^2$$

where we used the fact that the Gramm matrix $\{(\psi_i,\psi_j)_{\hat{T}}\}$ is spectrally equivalent to its diagonal $\{\|\psi_i\|_{\hat{T}}^2\}$. Then using the equality (cf. Lemma 1)

$$\|\phi\|_{,T}^2=\frac{|T|}{|\hat{T}|}\|\hat{\phi}\|_{0,\hat{T}}^2$$

we can derive the equivalence

$$\|v_T\|^2 \simeq \sum_{i \in I_T} \|v^i\|^2$$

with the same constants as on the reference simplex. After summation over all elements we get

$$\|v\|^2 = \sum_{T \in \mathcal{T}} \|v_T\|^2 \simeq \sum_{T \in \mathcal{T}} \sum_{i \in I_T} \|v^i\|^2 = \sum_{i=1}^{n} \|v^i\|^2$$

which verifies the second condition of the abstract theory. Then Theorems 8.1 and 8.2 from [8] imply that the scaled Jacobi and Gauss-Seidel smoothers satisfy the smoothing property. The symmetric Gauss-Seidel smoother can be analyzed by considering the sequence of spaces $\{\mathcal{V}^i\}_{i=1}^{2n}$ with $\mathcal{V}^i$ for $i = 1, \ldots, n$ as before and $\mathcal{V}^i = \mathcal{V}^{2n+1-i}$, $i = n+1, \ldots, 2n$. This choice results in doubling of the constant C_1, and C_2 can be taken to be the same. □

Our next step is to prove the approximation property (3.3) in the two cases we consider $\mathcal{V}_0 = \overline{\mathcal{V}}$ and $\mathcal{V}_0 = \mathcal{V}^c$. We consider the former case first.

Theorem 3. *Let $\overline{Q} : \mathcal{V} \to \overline{\mathcal{V}}$ be the L^2-orthogonal projection onto $\overline{\mathcal{V}}$. Then the following estimate holds*

$$\|v - \overline{Q}v\|^2 \leq Ch^2(\nabla v, \nabla v) \leq Ch^2 \mathcal{A}(v, v), \qquad \forall v \in \mathcal{V}.$$

Since $\lambda \simeq h^{-2}$ this is exactly the approximation property (3.3) *with $\widetilde{Q} = \overline{Q}$.*

Proof. Since the space $\overline{\mathcal{V}}$ is discontinuous the projection $\overline{Q}$ is local: $(\overline{Q}u)|_T$ is equal to the average of u over T

$$(\overline{Q}u)|_T = \frac{1}{|T|}(u, 1)_T.$$

Therefore $(v - \overline{Q}v)$ has zero average over every element T in $\mathcal{T}$ and we have the

estimate (cf. Theorem 3.1.4 in [14])

$$\|v - \overline{Q}v\|^2 = \sum_{T\in\mathcal{T}} \|v - \overline{Q}v\|^2_{0,T} \le Ch^2 \sum_{T\in\mathcal{T}} |v|^2_{1,T} = Ch^2(\nabla v, \nabla v).$$

Using the coercivity of $\mathcal{A}(\cdot,\cdot)$ in the norm (2.4) (see Proposition 2) we get

$$a_0(\nabla v, \nabla v) \le (a\nabla v, \nabla v) \le |||v|||^2 \le C\mathcal{A}(v,v)$$

where a_0 is the constant in the lower bound of (2.2). This completes the proof. □

Before we consider the case $\mathcal{V}_0 = \mathcal{V}^c$ we prove the following

Lemma 3. *Let $G = (V, E)$ be a connected graph with $V = \{1, \dots, n\}$ and let $E \subset V \times V$ be such that if $(i,j) \in E$ then $(j,i) \in E$ but $(i,i) \notin E$ for any $i \in V$. Let $v_i \in \mathbb{R}$, $i \in V$ then*

$$\sum_{i\in V} (v_i - \bar{v})^2 \le n^2 \sum_{\substack{(i,j)\in E \\ i<j}} (v_i - v_j)^2, \qquad \textit{where} \qquad \bar{v} = \frac{1}{n}\sum_{i\in V} v_i.$$

Proof. Let $k, l \in V$ be arbitrary, $k \ne l$, and let $(i_0, i_1, \dots, i_m)$ be a path connecting k and l where no vertex is repeated (so that $m < n$), i. e. $i_0 = k$, $i_m = l$ and $(i_{j-1}, i_j) \in E$, $j = 1, \dots, m$. Then

$$(v_k - v_l)^2 = \left(\sum_{j=1}^{m} (v_{i_{j-1}} - v_{i_j})\right)^2 \le m \sum_{j=1}^{m} (v_{i_{j-1}} - v_{i_j})^2 \le n \sum_{\substack{(i,j)\in E \\ i<j}} (v_i - v_j)^2. \tag{3.5}$$

We have

$$(v_k - \bar{v})^2 = \left(\sum_{l=1}^{n} \frac{1}{n}(v_k - v_l)\right)^2 \le n \sum_{l=1}^{n} \frac{1}{n^2}(v_k - v_l)^2 \le n \sum_{\substack{(i,j)\in E \\ i<j}} (v_i - v_j)^2$$

where in the last inequality we used (3.5) for each term $(v_k - v_l)^2$. Summation over $k \in V$ completes the proof. □

Now we consider the case $\mathcal{V}_0 = \mathcal{V}^c$.

Theorem 4. *There exists a projector $Q^c : \mathcal{V} \to \mathcal{V}^c$ such that*

$$\|v - Q^c v\|^2 \leq Ch \langle [\![v]\!], [\![v]\!] \rangle_{\mathcal{E}_i} \leq Ch^2 \mathcal{A}(v, v), \qquad \forall v \in \mathcal{V}.$$

Proof. Let let us denote by $\hat{\mathcal{N}}$ the following set of nodes on a reference d-simplex $\hat{T}$

$$\hat{\mathcal{N}} = \left\{ p \in \mathbb{R}^d : \hat{\lambda}_i(p) = k_i/r, k_i \in \{0, 1, \ldots, r\}, \forall i = 1, \ldots, d+1 \right\}$$

where $\{\hat{\lambda}_i\}_{i=1}^{d+1}$ are the barycentric functions on $\hat{T}$ which satisfy $\sum_{i=1}^{d+1} \hat{\lambda}_i = 1$. Then we define the set of all nodes for the triangulation $\mathcal{T}$ by

$$\mathcal{N} = \bigcup_{T \in \mathcal{T}} G_T(\hat{\mathcal{N}})$$

where G_T denotes the affine transformation from $\hat{T}$ to T. Note that with this choice of $\hat{\mathcal{N}}$ and the assumption that the mesh is regular (i. e. two elements either do not intersect or their intersection is a common vertex, edge, or face), we have that if an edge or face S is shared by two or more elements then the nodes on S from each of those elements coincide. For each node $\eta \in \mathcal{N}$ we denote the set of all elements that share that node by

$$\mathcal{T}_\eta = \{T \in \mathcal{T} : \eta \in \overline{T}\}.$$

Similarly, we denote the set of all *interior* faces that share a node η by

$$\mathcal{E}_\eta = \{F \in \mathcal{E}_i : \eta \in \overline{F}\}.$$

Note that the pair $(\mathcal{T}_\eta, \mathcal{E}_\eta)$ defines a graph for every node η in the following sense: the vertices of the graph are the elements in $\mathcal{T}_\eta$ and each face $F \in \mathcal{E}_\eta$ defines an edge of the graph connecting the two elements that share that face, $\mathcal{T}_F$ (note that $\mathcal{T}_F \subset \mathcal{T}_\eta$).

Given a function $v \in \mathcal{V}$, its degrees of freedom are given by (assuming the

elements T are open sets)

$$v_{\eta,T} = \lim_{\substack{x \longrightarrow \eta \\ x \in T}} v(x), \qquad \forall (\eta, T) \in \mathcal{N} \times \mathcal{T} : \eta \in \overline{T}.$$

Note that $v \in \mathcal{V}^c$ if and only if $\forall \eta \in \mathcal{N}$ we have

$$v_{\eta,T_1} = v_{\eta,T_2} = v_\eta, \qquad \forall T_1, T_2 \in \mathcal{T}_\eta.$$

Now we can define the projector $Q^c : \mathcal{V} \to \mathcal{V}^c$ by defining its values at the nodes

$$(Q^c v)(\eta) = \frac{1}{|\mathcal{T}_\eta|} \sum_{T \in \mathcal{T}_\eta} v_{\eta,T}, \qquad \forall \eta \in \mathcal{N},$$

where $|\mathcal{T}_\eta|$ stands for the number of elements in the set $\mathcal{T}_\eta$. As a first step in proving the estimate for Q^c we will find estimates for $\|v\|^2$ and $\langle [\![v]\!], [\![v]\!] \rangle_{\mathcal{E}_i}$ in terms of the degrees of freedom $\{v_{\eta,T}\}$. By mapping the polynomial $v|_T$ to a reference simplex $\hat{T}$ we get (cf. (2.5))

$$\|v\|^2_{0,T} \simeq h^d \|\hat{v}\|^2_{0,\hat{T}} = h^d \sum_{i \in \hat{\mathcal{N}}} \sum_{j \in \hat{\mathcal{N}}} \hat{v}(i)\hat{v}(j)(\hat{\psi}_i, \hat{\psi}_j)_{\hat{T}} \simeq h^d \sum_{i \in \hat{\mathcal{N}}} \hat{v}(i)^2,$$

where $\hat{\psi}_i$ denotes the nodal basis function corresponding to the node i; for the last equivalence we used the fact that the Gramm matrix $\{(\hat{\psi}_i, \hat{\psi}_j)_{\hat{T}}\}$ is spectrally equivalent to the identity. Since $\hat{v}(\hat{\eta}) = v_{\eta,T}$, where $\eta = G_T(\hat{\eta})$ we have

$$\|v\|^2 = \sum_{T \in \mathcal{T}} \|v\|^2_{0,T} \simeq h^d \sum_{T \in \mathcal{T}} \sum_{\eta \in \mathcal{N} \cap \overline{T}} v^2_{\eta,T} = h^d \sum_{\eta \in \mathcal{N}} \sum_{T \in \mathcal{T}_\eta} v^2_{\eta,T}$$

Let us introduce the notation

$$[\![v]\!]_{\eta,F} = v_{\eta,T_1} - v_{\eta,T_2}, \qquad \forall (\eta, F) \in \mathcal{N} \times \mathcal{E}_i : \eta \in \overline{F},$$

where T_1 and T_2 are the two elements that have F as a common face and the normal vector $\mathbf{n}$ to F points outside of T_1 (this determines the sign of $[\![v]\!]_{\eta,F}$). Since $[\![v]\!]|_F$ is

a polynomial of degree r, using an argument similar to above one can show that

$$\langle [\![v]\!], [\![v]\!] \rangle_{\mathcal{E}_i} \simeq h^{d-1} \sum_{F \in \mathcal{E}_i} \sum_{\eta \in \mathcal{N} \cap \overline{F}} [\![v]\!]^2_{\eta,F} = h^{d-1} \sum_{\eta \in \mathcal{N}} \sum_{F \in \mathcal{E}_\eta} [\![v]\!]^2_{\eta,F}.$$

Let $v \in \mathcal{V}$, $v^c = Q^c v$, and denote $v^c_\eta = v^c(\eta)$, $\forall \eta \in \mathcal{N}$ then

$$\|v - Q^c v\|^2 \simeq h^d \sum_{\eta \in \mathcal{N}} \sum_{T \in \mathcal{T}_\eta} \left(v_{\eta,T} - v^c_\eta \right)^2 = h^d \sum_{\eta \in \mathcal{N}_i} \sum_{T \in \mathcal{T}_\eta} \left(v_{\eta,T} - v^c_\eta \right)^2,$$

where $\mathcal{N}_i$ is the set of all nodes which belong to at least two elements; for all other nodes $v_{\eta,T} = v^c_\eta$. It is clear that if $\eta \in \mathcal{N} \setminus \mathcal{N}_i$ then the set $\mathcal{E}_\eta$ is empty because $\mathcal{T}_\eta$ has only one element and consequently

$$\langle [\![v]\!], [\![v]\!] \rangle_{\mathcal{E}_i} \simeq h^{d-1} \sum_{\eta \in \mathcal{N}_i} \sum_{F \in \mathcal{E}_\eta} [\![v]\!]^2_{\eta,F}.$$

It is now clear that the following estimate (with C independent of η and h)

$$\sum_{T \in \mathcal{T}_\eta} \left(v_{\eta,T} - v^c_\eta \right)^2 \leq C \sum_{F \in \mathcal{E}_\eta} [\![v]\!]^2_{\eta,F}, \qquad \forall \eta \in \mathcal{N}_i \tag{3.6}$$

will prove the first estimate of the theorem. To prove such an estimate we will use Lemma 3 but first we need to see that the graph $(\mathcal{T}_\eta, \mathcal{E}_\eta)$ is connected. This is not hard to see when N is in the interior of Ω. When $\eta \in \partial\Omega$ one can use the assumption that the boundary is Lipschitz. Thus, we can apply Lemma 3 to prove (3.6) with $C = |\mathcal{T}_\eta|^2$ which can be bounded independent of η and h because of the assumptions on the mesh $\mathcal{T}$. The proof of the estimate

$$\|v - Q^c v\|^2 \leq Ch \langle [\![v]\!], [\![v]\!] \rangle_{\mathcal{E}_i}$$

is now completed. Using the definition (2.4) of the norm $|||\cdot|||$ and the coercivity of $\mathcal{A}(\cdot,\cdot)$ with respect to it we get

$$Ch \langle [\![v]\!], [\![v]\!] \rangle_{\mathcal{E}_i} \leq Ch^2 \left\langle \kappa h_{\mathcal{E}}^{-1} a_{\mathcal{E}} [\![v]\!], [\![v]\!] \right\rangle_{\mathcal{E}_i \cup \mathcal{E}_D} \leq Ch^2 |||v|||^2 \leq Ch^2 \mathcal{A}(v,v),$$

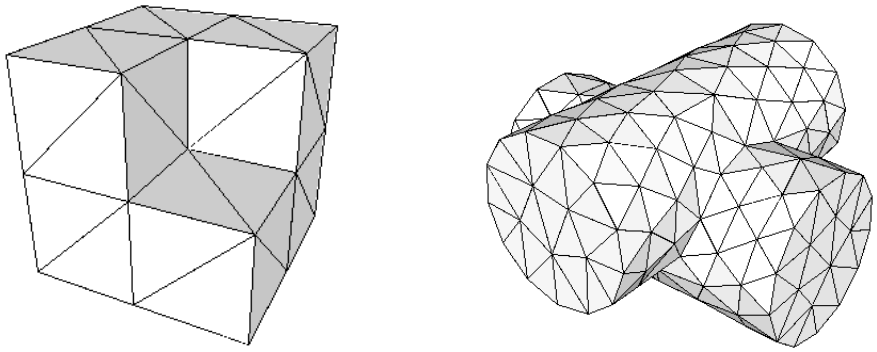

Fig. 3.1. Coarse meshes for the second (left) and third (right) test problems.

which finalizes the proof of the theorem. □

3.1.3. Numerical Experiments

We present numerical results for three test problems of elliptic equations with homogeneous Dirichlet boundary conditions on the whole boundary:

- *Test Problem 1*: The equation $-\Delta u = 1$, $u|_{\partial\Omega} = 0$ in the cube $\Omega = (0,1)^3$.
- *Test Problem 2*: The equation $-\nabla \cdot (a \nabla u) = 1$, $u|_{\partial\Omega} = 0$ in $\Omega = (0,1)^3 \setminus [0.5,1)^3$ (see Figure 3.1) where the coefficient a has jumps (a 3-D checkerboard pattern) as follows: $a = 1$, in $(I_1 \times I_1 \times I_1) \cup (I_2 \times I_2 \times I_1) \cup (I_1 \times I_2 \times I_2) \cup (I_2 \times I_1 \times I_2)$ and $a = \epsilon$, in the other parts of Ω, where $I_1 = (0, 0.5]$ and $I_2 = (0.5, 1]$, and we vary the value of ϵ to test the efficiency of the preconditioners with respect to the size of the jumps.
- *Test Problem 3*: The equation $-\Delta u = 1$, $u|_{\partial\Omega} = 0$ in the domain shown on Figure 3.1. The discretization of the domain (shown on the figure) is unstructured and has some thin elements, i. e. elements T for which ρ_T / h_T is small.

For all test examples we have used a coarse tetrahedral mesh (corresponding to "Level 0") which is uniformly refined to form a sequence of meshes with "Level k" denoting

the mesh obtained after k uniform refinements. The discretization is using linear and quadratic elements. The value of the penalty term was experimentally chosen to be $\kappa = 10$ for linear, and $\kappa = 20$ for quadratic finite elements for Test Problems 1, 2; for Test Problem 3 we had to increase κ because of thin elements in the mesh, namely we used $\kappa = 15$ for linear and $\kappa = 30$ for quadratic FE. We run experiments with both two-level preconditioners presented above:

- Method I, using the coarse spaces $\mathcal{V}^c$, and
- Method II, using the coarse space $\overline{\mathcal{V}}$.

In all cases we use one pre-smoothing and one post-smoothing step with a symmetric Gauss-Seidel smoother. The preconditioned conjugate gradient (PCG) method was used to approximately solve the resulting linear systems with relative accuracy of 10^{-8}, i. e. we iterate until the error measured by $(r^t Br)^{1/2}$ is reduced by a factor of 10^8. The coarse level systems that need to be solved to apply the preconditioner are also solved by a PCG iteration with the same relative accuracy using symmetric Gauss-Seidel as a preconditioner.

We report the number of iterations (denoted by "iter" in the tables) it took the PCG method to converge and the average reduction factor (denoted by "arf") which is defined as

$$\text{arf} = \left(\frac{(r_n^t B r_n)^{1/2}}{(r_0^t B r_0)^{1/2}} \right)^{1/n}$$

where n is the number of iterations and r_i is the i-th residual. This means that $(\text{arf})^n$ will be roughly 10^8 but it gives a finer way of measuring convergence compared to the number of iterations.

The results for Test Problem I using linear and quadratic finite elements (FE) are presented in Tables 3.1 and 3.2, respectively. The columns in the tables represent the refinement level. The rows labeled with "$\mathcal{S}$ dof" give the number of degrees of

freedom (dof) in the space $\mathcal{S}$ which is one of the spaces $\mathcal{V}$, $\mathcal{V}^c$, or $\overline{\mathcal{V}}$. In the other rows (labeled with "iter/arf") we give the number of iterations and the average reduction factors when using the two-level preconditioner based on the corresponding coarse space.

Table 3.1. Two-level preconditioners, Test Problem 1, linear FE

	Level 2	Level 3	Level 4	Level 5	Level 6
$\mathcal{V}$ dof	3,072	24,576	196,608	1,572,864	12,582,912
$\mathcal{V}^c$ dof	189	1,241	9,009	68,705	536,769
$\mathcal{V}^c$, iter/arf	13/0.2307	13/0.2304	13/0.2238	12/0.2094	12/0.1973
$\overline{\mathcal{V}}$ dof	768	6,144	49,152	393,216	3,145,728
$\overline{\mathcal{V}}$, iter/arf	19/0.3786	21/0.4047	21/0.4028	21/0.3999	20/0.3967

Table 3.2. Two-level preconditioners, Test Problem 1, quadratic FE

	Level 1	Level 2	Level 3	Level 4	Level 5
$\mathcal{V}$ dof	960	7,680	61,440	491,520	3,932,160
$\mathcal{V}^c$ dof	189	1,241	9,009	68,705	536,769
$\mathcal{V}^c$, iter/arf	9/0.1237	10/0.1478	10/0.1425	9/0.1267	9/0.1151
$\overline{\mathcal{V}}$ dof	96	768	6,144	49,152	393,216
$\overline{\mathcal{V}}$, iter/arf	18/0.3591	27/0.5027	31/0.5425	30/0.5408	30/0.5353

From the tables we see that the number of iterations remains bounded when we refine the mesh. This result agrees with our theoretical results which say that the condition number of the preconditioned system is bounded which implies that the number of iterations is also bounded. When linear elements are used Method II uses roughly two times more iterations than Method I even though $\mathcal{V}^c$ (the coarse space

of Method I) has less degrees of freedom than $\overline{\mathcal{V}}$ (the coarse space of Method II); on the other hand, note that the matrix on $\overline{\mathcal{V}}$ has very simple sparsity pattern — it has at most 5 non-zero entries per row. When quadratic elements are used Method I uses about three times less iterations and has just slightly larger number of dofs in the coarse space compared to Method II. Therefore we can say that Method I is better than Method II for this Test Problem.

The results for Test Problem 2 using linear FE are given in Table 3.3 where we give the number of iterations and average reduction factors for Method I and Method II in the top and bottom parts of the table, respectively.

For both methods, if we consider a fixed value of ϵ, the number of iterations slowly increases with the first 2-3 levels of refinement and then stabilizes (it is almost constant) for the finer triangulations. For Method I, the number of iterations at which this stabilization occurs doubles when ϵ decreases from 1 to 10^{-3} and then slowly decreases with ϵ. This can be clearly seen in the last column ("Level 5") in the top part of the table. However, such dependence is fairly weak for such wide range of the parameter ϵ. For Method II, the number of iterations is almost completely independent of ϵ. Notice that even the average reduction factors seem to converge to a fixed number as $\epsilon \to 0$. Comparing the results for both methods, we can see that Method II converges in less iterations than Method I for the smaller values of ϵ when the mesh is fine enough. This suggests that Method II is probably the better choice for problems with large jumps in the elliptic coefficient.

In our analysis of the two-level Methods I and II, many of estimates we proved depend on the ratio a_1/a_0 where a_0 and a_1 are the constants from assumption (2.2). For Test Problem 2, $a_1/a_0 = \epsilon$ and therefore our theory is not independent of ϵ. However, the presented numerical results suggest that the convergence is independent of ϵ.

Table 3.3. Two-level preconditioners, Test Problem 2, linear FE

	Level 1	Level 2	Level 3	Level 4	Level 5
$\mathcal{V}$ dof	1,344	10,752	86,016	688,128	5,505,024
$\mathcal{V}^c$ dof	117	665	4,401	31,841	241,857
$\epsilon = 1$	14/0.2432	14/0.2605	14/0.2545	13/0.2412	13/0.2313
$\epsilon = 0.1$	15/0.2827	18/0.3450	19/0.3711	19/0.3745	19/0.3700
$\epsilon = 0.01$	17/0.3374	21/0.4001	25/0.4720	27/0.4990	27/0.5055
$\epsilon = 0.001$	17/0.3325	21/0.4054	26/0.4798	28/0.5106	29/0.5252
$\epsilon = 10^{-4}$	17/0.3196	21/0.4022	25/0.4768	27/0.5037	28/0.5128
$\epsilon = 10^{-5}$	16/0.3012	20/0.3975	24/0.4540	26/0.4843	27/0.4984
$\epsilon = 10^{-6}$	15/0.2909	19/0.3791	22/0.4275	24/0.4622	25/0.4743
$\overline{\mathcal{V}}$ dof	336	2,688	21,504	172,032	1,376,256
$\epsilon = 1$	18/0.3527	20/0.3950	21/0.4051	21/0.4024	21/0.3980
$\epsilon = 0.1$	20/0.3734	21/0.4036	21/0.4099	21/0.4098	21/0.4069
$\epsilon = 0.01$	19/0.3644	20/0.3938	21/0.4017	21/0.4048	21/0.4039
$\epsilon = 0.001$	18/0.3552	20/0.3917	21/0.4002	21/0.4055	21/0.4042
$\epsilon = 10^{-4}$	18/0.3467	20/0.3883	21/0.3997	21/0.4056	21/0.4043
$\epsilon = 10^{-5}$	17/0.3374	19/0.3716	21/0.3997	21/0.4056	21/0.4043
$\epsilon = 10^{-6}$	17/0.3314	19/0.3674	21/0.3997	21/0.4056	21/0.4043

Table 3.4. Two-level preconditioners, Test Problem 3, linear FE

	Level 1	Level 2	Level 3	Level 4
$\mathcal{V}$ dof	24,032	192,256	1,538,048	12,304,384
$\mathcal{V}^c$ dof	1,445	9,693	70,633	538,513
$\mathcal{V}^c$, iter/arf	18/0.3511	18/0.3485	17/0.3359	19/0.3739
$\overline{\mathcal{V}}$ dof	6,008	48,064	384,512	3,076,096
$\overline{\mathcal{V}}$, iter/arf	34/0.5743	37/0.6077	40/0.6267	43/0.6494

In Table 3.4 we present the results for Test Problem 3. This problem is a test for the preconditioners on unstructured mesh having thin elements. Such elements introduce geometrical anisotropy which makes the problem harder. For the mesh sequence that we consider this anisotropy is not very strong — the highest aspect ratio is around 12. The numerical results show that the number of iterations is almost independent of the refinement level. However, if we compare the results for this Test Problem and Test Problem 1 we can see an increase in the number of iterations. This increase is slightly larger for Method II than it is for Method I. From this we can conclude that both two-level methods are sensitive to geometrical anisotropies which is not unexpected since we use a simple point smoother.

3.2. Multilevel Methods

3.2.1. Multigrid Setup and Algorithms

We assume that we have a sequence of nested simplicial triangulations (we will also call them meshes) of the domain Ω which we denote with $\mathcal{T}_k$, $k = 1, \ldots, J$, with $\mathcal{T}_1$ being the coarsest triangulation. The triangulations are nested in the sense that every element in $\mathcal{T}_k$ is the union of elements in the finer mesh $\mathcal{T}_{k+1}$. We will also assume that all meshes are regular, i. e. any two elements (in the same mesh) either do not intersect or their intersection is a common vertex, edge, or face. Finally, we assume that the elements are shape regular and the meshes are globally quasi-uniform, that is there exist constants $\gamma > 0$ and $c > 0$ such that

$$\frac{h_T}{\rho_T} \leq \gamma, \qquad \forall T \in \mathcal{T}_k, \quad \forall k = 1, \ldots, J \qquad \text{(shape regularity)}$$

$$ch_k \leq h_T \leq h_k, \qquad \forall T \in \mathcal{T}_k, \quad \forall k = 1, \ldots, J \qquad \text{(global quasi-uniformity)}$$

where ρ_T denotes the diameter of the largest ball contained in T, h_T is the diameter of T, and

$$h_k = \max_{T \in \mathcal{T}_k} h_T, \qquad k = 1, \ldots, J.$$

Similarly to the notation introduced in the previous chapter we will use $\mathcal{E}_k$, $\mathcal{E}_i^k$, $\mathcal{E}_b^k$, $\mathcal{E}_D^k$, $\mathcal{E}_N^k$ to denote the different sets of faces with the index k indicating that the sets are faces of the mesh $\mathcal{T}_k$. To define $\mathcal{E}_D^k$ we need to assume that Γ_D is the union of some boundary faces on the coarsest level, $\mathcal{E}_b^0$, and as a consequence — on all levels. We will also use the "broken" Sobolev spaces

$$H^s(\mathcal{T}_k) = \left\{ v \in L^2(\Omega) : v|_T \in H^s(T), \ \forall T \in \mathcal{T}_k \right\}, \text{ for } s \geq 0,$$

the discrete spaces of discontinuous piecewise polynomial functions of degree $r \geq 1$:

$$\mathcal{V}_k = \left\{ v \in L^2(\Omega) : v|_T \in P_r(T),\ \forall T \in \mathcal{T}_k \right\},$$

and the corresponding continuous and piecewise constant discrete spaces

$$\mathcal{V}_k^c = \left\{ v \in \mathcal{C}(\overline{\Omega}) : v|_T \in P_r(T),\ \forall T \in \mathcal{T}_k \right\} = \mathcal{C}(\overline{\Omega}) \cap \mathcal{V}_k$$
$$\overline{\mathcal{V}}_k = \left\{ v \in L^2(\Omega) : v|_T = \text{const},\ \forall T \in \mathcal{T}_k \right\}.$$

For functions u and v in $H^s(\mathcal{T}_k)$, $s > \frac{3}{2}$, the k-th level symmetric interior penalty (SIPG) bilinear form is given by

$$\begin{aligned}\mathcal{A}_k(u,v) = (a\nabla u, \nabla v) &- \langle \{a\nabla u \cdot \mathbf{n}_k\}, [\![v]\!] \rangle_{\mathcal{E}_i^k \cup \mathcal{E}_D^k} \\ &- \langle \{a\nabla v \cdot \mathbf{n}_k\}, [\![u]\!] \rangle_{\mathcal{E}_i^k \cup \mathcal{E}_D^k} + \left\langle \kappa h_{\mathcal{E}_k}^{-1} a_{\mathcal{E}_k} [\![u]\!], [\![v]\!] \right\rangle_{\mathcal{E}_i^k \cup \mathcal{E}_D^k}\end{aligned}$$

and the corresponding linear form by

$$\mathcal{L}_k(v) = (f, v) + \langle u_N, v \rangle_{\mathcal{E}_N^k} - \langle u_D, a\nabla v \cdot \mathbf{n}_k \rangle_{\mathcal{E}_D^k} + \left\langle \kappa h_{\mathcal{E}_k}^{-1} a_{\mathcal{E}_k} u_D, v \right\rangle_{\mathcal{E}_D^k}$$

where once again we use ∇u to denote the element-by-element derivative if u. With these definitions, the interior penalty discontinuous Galerkin discretization method for our elliptic problem (2.1) reads: find $u_h \in \mathcal{V}_J$ such that

$$\mathcal{A}_J(u_h, v) = \mathcal{L}_J(v), \quad \forall v \in \mathcal{V}_J. \tag{3.7}$$

For convenience, we will use $|||\cdot|||_k$ to denote the energy norm on $\mathcal{V}_k$:

$$|||u|||_k = \mathcal{A}_k(u,u)^{\frac{1}{2}}, \qquad \forall u \in \mathcal{V}_k$$

and the k-th level $|||\cdot|||$ from (2.4) we will now denote by

$$|||v|||_{*,k}^2 = (a\nabla v, \nabla v) + \left\langle \kappa h_{\mathcal{E}_k}^{-1} a_{\mathcal{E}_k} [\![v]\!], [\![v]\!] \right\rangle_{\mathcal{E}_i^k \cup \mathcal{E}_D^k}, \qquad \forall v \in H^1(\mathcal{T}_k).$$

As we showed in Proposition 2, for large enough κ, we have the following norm equivalence on $\mathcal{V}_k$:

$$|||v|||_k \simeq |||v|||_{*,k}\,, \qquad \forall v \in \mathcal{V}_k, \tag{3.8}$$

with constants independent of h_k.

We will now define a general multigrid algorithm based on a sequence of nested spaces with bilinear forms that are not inherited. The algorithm presented here is a version of the one given in Section 7 of [8] for the case of nested spaces.

Let $M_1 \subset M_2 \subset \cdots \subset M_{\widetilde{J}} \subset L^2(\Omega)$ be sequence of nested finite dimensional spaces and let $A_k : M_k \times M_k \to \mathbb{R}$ be given s. p. d. bilinear forms. Define the operators $A_k : M_k \to M_k$, $Q_k : L^2(\Omega) \to M_k$, and $P_k : M_{k+1} \to M_k$ by

$$\begin{aligned}
(A_k u, v) &= A_k(u, v), && \forall v \in M_k, \quad k = 1, \ldots, \widetilde{J} \\
(Q_k u, v) &= (u, v), && \forall v \in M_k, \quad k = 1, \ldots, \widetilde{J} \\
A_k(P_k u, v) &= A_{k+1}(u, v), && \forall v \in M_k, \quad k = 1, \ldots, \widetilde{J} - 1
\end{aligned}$$

where $(\cdot, \cdot)$ denotes the inner product in $L^2(\Omega)$. Assume we are given the smoothing operators $R_k : M_k \to M_k$ and set

$$R_k^{(\ell)} = \begin{cases} R_k & \text{if } \ell \text{ is odd,} \\ R_k^t & \text{if } \ell \text{ is even,} \end{cases}$$

where R_k^t denotes the adjoint of R_k with respect to $(\cdot, \cdot)$. Let m_k be a given number of pre- and post-smoothing iterations, p — a given number of correction steps, and set $B_1 = A_1^{-1}$, then B_k is defined recursively: given $g \in M_k$

1. pre-smoothing: define $x^\ell \in M_k$, $\ell = 0, \ldots, m_k$ by: set $x^0 = 0$ and

$$x^\ell = x^{\ell-1} + R_k^{(\ell+m_k)}(g - A_k x^{\ell-1}), \quad \ell = 1, \ldots, m_k.$$

2. correction: define $y^{m_k} = x^{m_k} + q^p$ where $q^p \in M_{k-1}$ is defined by: set $q^0 = 0$ and then for $\ell = 1, \ldots, p$ set

$$q^\ell = q^{\ell-1} + B_{k-1}[Q_{k-1}(g - A_k x^{m_k}) - A_{k-1} q^{\ell-1}].$$

3. post-smoothing: define $y^\ell \in M_k$, $\ell = m_k + 1, \ldots, 2m_k$ by

$$y^\ell = y^{\ell-1} + R_k^{(\ell+m_k)}(g - A_k y^{\ell-1}).$$

4. $B_k g = y^{2m_k}$.

We will consider the multigrid algorithms arising from the following two choices of spaces M_k and bilinear forms A_k:

1. Method I, based on the spaces $\mathcal{V}_k^c$: $\widetilde{J} = J + 1$ and

$$M_k = \begin{cases} \mathcal{V}_k^c, & k = 1, \ldots, J \\ \mathcal{V}_J, & k = J + 1 \end{cases} \qquad A_k = \begin{cases} \mathcal{A}_k, & k = 1, \ldots, J \\ \mathcal{A}_J, & k = J + 1. \end{cases}$$

2. Method II, based on the spaces $\overline{\mathcal{V}}_k$: $\widetilde{J} = J + 1$ and

$$M_k = \begin{cases} \overline{\mathcal{V}}_k, & k = 1, \ldots, J \\ \mathcal{V}_J, & k = J + 1 \end{cases} \qquad A_k = \begin{cases} \mathcal{A}_k, & k = 1, \ldots, J \\ \mathcal{A}_J, & k = J + 1. \end{cases}$$

Remark 2. *One can generalize the above two methods by choosing an integer* $j_0 \in [1, J]$ *and setting*

$$M_k = \begin{cases} \mathcal{V}_k^c, & k = 1, \ldots, j_0 \\ \mathcal{V}_{k-1}, & k = j_0 + 1, \ldots, J + 1 \end{cases} \qquad A_k = \begin{cases} \mathcal{A}_k, & k = 1, \ldots, j_0 \\ \mathcal{A}_{k-1}, & k = j_0 + 1, \ldots, J + 1 \end{cases}$$

and similarly using the spaces $\overline{\mathcal{V}}_k$.

Remark 3. *A natural choice is to let* $\widetilde{J} = J$, $M_k = \mathcal{V}_k$, *and* $A_k = \mathcal{A}_k$, $\forall k$. *The resulting algorithm was considered and analyzed by Gopalakrishnan and Kanschat in their paper [19]. Our analysis of Method I in the following section was strongly influenced by their work.*

Remark 4. *Another approach is to choose* $A_k = \mathcal{A}_J$, $\forall k$, *in all of the above mentioned methods. This leads to a nested-inherited setting. However, the penalty term in* $\mathcal{A}_J$ *which is preserved in the coarse levels operators introduces high frequencies in them which in turn makes the standard approach to the multigrid analysis inapplicable. A version of this method will be used in our algebraic approach presented in the next chapter.*

3.2.2. Analysis

In this section we analyze Method I using the approach from [19] which is based on the abstract theory from [8]. We begin with the following error estimate:

Lemma 4. *Consider the case of homogeneous Dirichlet boundary condition,* $u_D = 0$, *and assume that the solution* U *of (2.1) is in* $H^{1+\alpha}(\Omega)$ *for some* $\frac{1}{2} < \alpha \leq 1$. *Let* $U_k \in \mathcal{V}_k$ *(or* $\mathcal{V}_k^c$, *we will consider both cases simultaneously) be the solution of*

$$\mathcal{A}_k(U_k, v) = \mathcal{L}_k(v), \quad \forall v \in \mathcal{V}_k \ (\mathcal{V}_k^c).$$

Then the following error estimate holds

$$|||U - U_k|||_{*,k} \leq C h_k^{\alpha} \|U\|_{1+\alpha}$$

with a constant C *independent of* h_k.

Proof. As we showed in Proposition 1 that the IP method is consistent:

$$\mathcal{A}_k(U, v) = \mathcal{L}_k(v), \quad \forall v \in H^{1+\alpha}(\mathcal{T}_k)$$

which combined with the definition of U_k gives the Galerkin orthogonality:

$$\mathcal{A}_k(U - U_k, v) = 0, \quad \forall v \in \mathcal{V}_k \ (\mathcal{V}_k^c). \tag{3.9}$$

We will use the following norm on $H^{1+\alpha}(\mathcal{T}_k)$

$$|||u|||_{\alpha,k}^2 = |||u|||_{*,k}^2 + \sum_{T\in\mathcal{T}_k} h_k^{2\alpha}|u|_{1+\alpha,T}^2$$

which is equivalent to $|||\cdot|||_{*,k}$ on $\mathcal{V}_k$ due to the inverse inequality $|u|_{1+\alpha,T} \le Ch_k^{-\alpha}|u|_{1,T}$. We want to show that the bilinear form $\mathcal{A}_k(\cdot,\cdot)$ is bounded in the norm $|||\cdot|||_{\alpha,k}$ for arbitrary functions in $H^{1+\alpha}(\mathcal{T}_k)$. Let $T \in \mathcal{T}_k$ be an element and F one of its faces and let $\phi \in H^\alpha(T)$ then using a trace theorem on a reference d-simplex $\hat{T}$ we have

$$\|\hat{\phi}\|_{0,\hat{F}}^2 \le C\|\hat{\phi}\|_{\alpha,\hat{T}}^2, \quad \text{with } \hat{F} = G^{-1}(F), \text{ and } \hat{\phi}(\hat{x}) = \phi(G(\hat{x})),$$

where $G : \hat{T} \to T$ is the affine transformation from $\hat{T}$ to T: $x = G(\hat{x}) = B\hat{x} + b$. Using the definition of the seminorm $|\cdot|_\alpha$ for $\alpha < 1$ and change of the variables we have

$$|\hat{\phi}|_{\alpha,\hat{T}}^2 = \int_{\hat{T}}\int_{\hat{T}} \frac{|\hat{\phi}(\hat{x}) - \hat{\phi}(\hat{y})|^2}{|\hat{x} - \hat{y}|^{d+2\alpha}}\, d\hat{x}\, d\hat{y} \le \frac{\|B\|^{d+2\alpha}}{|\det B|^2}\int_T\int_T \frac{|\phi(x) - \phi(y)|^2}{|x - y|^{d+2\alpha}}\, dx\, dy$$

where we used the equality $dx = |\det B| d\hat{x}$ and the estimate

$$|x - y| = |B\hat{x} + b - B\hat{y} - b| \le \|B\||\hat{x} - \hat{y}|.$$

Since $\|B\| \lesssim h_k$ and $|\det B| \simeq h_k^d$ (see [14], Section 3.1), we get

$$|\hat{\phi}|_{\alpha,\hat{T}}^2 \lesssim h_k^{2\alpha-d}|\phi|_{\alpha,T}^2$$

which is also valid for $\alpha = 1$ (see [14], Section 3.1). Using this estimate, the trace

inequality from above, and the equalities (2.5) on page 13 we obtain the estimate

$$\|\phi\|_{0,F}^2 \simeq h_k^{d-1}\|\hat{\phi}\|_{0,\hat{F}}^2 \lesssim h_k^{d-1}\left(\|\hat{\phi}\|_{0,\hat{T}}^2 + |\hat{\phi}|_{\alpha,\hat{T}}^2\right) \lesssim h_k^{-1}\left(\|\phi\|_{0,T}^2 + h_k^{2\alpha}|\phi|_{\alpha,T}^2\right).$$

Let $u, v \in H^{1+\alpha}(\mathcal{T}_k)$ then using the above inequality for the derivatives of u one easily shows that

$$\int_F (a\nabla u \cdot \mathbf{n})\,[\![v]\!] \le \|a\nabla u \cdot \mathbf{n}\|_{0,F}\|\,[\![v]\!]\|_{0,F} \le C\left(|u|_{1,T}^2 + h_k^{2\alpha}|u|_{1+\alpha,T}^2\right)^{\frac{1}{2}} h_{\mathcal{E}_k}^{-\frac{1}{2}}\|\,[\![v]\!]\|_{0,F}$$

which, in turn, implies that

$$\begin{aligned}\langle\{a\nabla u \cdot \mathbf{n}\}, [\![v]\!]\rangle_{\mathcal{E}_i^k \cup \mathcal{E}_D^k} &\le C\left(\sum_{T\in\mathcal{T}_k} |u|_{1,T}^2 + h_k^{2\alpha}|u|_{1+\alpha,T}^2\right)^{\frac{1}{2}} \left\langle h_{\mathcal{E}_k}^{-1}[\![v]\!], [\![v]\!]\right\rangle_{\mathcal{E}_i^k\cup\mathcal{E}_D^k}^{\frac{1}{2}} \\ &\le C\,|||u|||_{\alpha,k}\,|||v|||_{\alpha,k}.\end{aligned}$$

This estimate implies the boundedness of $\mathcal{A}_k(\cdot,\cdot)$ in the norm $|||\cdot|||_{\alpha,k}$. Let $w_k \in \mathcal{V}_k$ ($\mathcal{V}_k^c$) then using the coercivity part of (3.8), the Galerkin orthogonality (3.9) and the boundedness we just established we get

$$\begin{aligned}|||U_k - w_k|||_{\alpha,k}^2 &\le C\,|||U_k - w_k|||_{*,k}^2 \le C\mathcal{A}_k(U_k - w_k, U_k - w_k) \\ &= C\mathcal{A}_k(U - w_k, U_k - w_k) \le C\,|||U - w_k|||_{\alpha,k}\,|||U_k - w_k|||_{\alpha,k}\end{aligned}$$

which simplifies to

$$|||U_k - w_k|||_{\alpha,k} \le C\,|||U - w_k|||_{\alpha,k}.$$

Combining this with the triangle inequality gives

$$|||U - U_k|||_{\alpha,k} \le |||U - w_k|||_{\alpha,k} + |||U_k - w_k|||_{\alpha,k} \le C\,|||U - w_k|||_{\alpha,k}$$

that is U_k is a quasi-optimal approximation to U from the space $\mathcal{V}_k$ ($\mathcal{V}_k^c$) in the norm $|||\cdot|||_{\alpha,k}$. We take $w_k = \Pi_k U$ to be the Scott-Zhang interpolation of U in the continuous piecewise linear space which preserves the homogeneous boundary condition (see [24])

so that we have

$$[\![U - w_k]\!]|_F = 0, \qquad \forall F \in \mathcal{E}_i^k \cup \mathcal{E}_D^k.$$

Note that we take w_k to be linear even when $r > 1$, i. e. when the space $\mathcal{V}_k$ ($\mathcal{V}_k^c$) uses higher degree polynomials. Using the the error estimate

$$|U - w_k|_{1,\Omega} \leq C h_k^{\alpha} \|U\|_{1+\alpha,\Omega}$$

and the equality (here we use the linearity of $w_k|_T$)

$$|U - w_k|_{1+\alpha,T} = |U|_{1+\alpha,T},$$

we obtain

$$\begin{aligned} |||U - U_k|||_{\alpha,k}^2 &\leq C\, |||U - w_k|||_{\alpha,k}^2 \leq C|U - w_k|_{1,\Omega}^2 + C \sum_{T \in \mathcal{T}_k} h_k^{2\alpha} |U - w_k|_{1+\alpha,T}^2 \\ &\leq C h_k^{2\alpha} \|U\|_{1+\alpha,\Omega}^2. \end{aligned}$$

This completes the proof. □

Proposition 4. *Let λ_k denote the largest eigenvalue of A_k. Then*

$$\lambda_k \equiv \sup_{v \in M_k} \frac{(A_k v, v)}{(v, v)} \simeq h_k^{-2}$$

where $h_{J+1} = h_J$ (when $k = J + 1$).

Proof. For $k = J+1$, $M_{J+1} = \mathcal{V}_J$, this is exactly Proposition 3. For $k \leq J$, $M_k = \mathcal{V}_k^c$, the estimate $\lambda_k \leq C h_k^{-2}$ follows from the same Proposition and the estimate from below can be easily obtained from the estimates therein, considering a continuous nodal basis function in $\mathcal{V}_k^c$. □

Remark 5. *Using the estimates from [23] (see page 454 and Theorem 3.1), one can*

show that the smallest eigenvalue of A_k is independent of h_k:

$$\inf_{v \in M_k} \frac{(A_k v, v)}{(v, v)} \simeq 1$$

and therefore the condition number of A_k is $\mathcal{O}(h_k^{-2})$.

The next theorem is a reformulation of Theorem 7.4 form [8].

Theorem 5. *Assume that the following two conditions are satisfied:*

1. *There exists $\omega > 0$ not depending on k such that*

$$\left(\frac{\omega}{\lambda_k}\right) \|v\|^2 \leq (\overline{R}_k v, v), \quad \forall v \in M_k, \ k = 2, \ldots, J+1 \tag{3.10}$$

   where $\overline{R}_k = R_k + R_k^t + R_k^t A_k R_k$.

2. *For some α with $0 < \alpha \leq 1$ there exists C_P independent of k such that*

$$|A_k((I - P_{k-1})v, v)| \leq C_P \left(\frac{\|A_k v\|^2}{\lambda_k}\right)^{\alpha} [A_k(v, v)]^{1-\alpha}, \tag{3.11}$$

   for all $v \in M_k$, $k = 2, \ldots, J+1$.

Assume also that for some $1 < \beta_0 \leq \beta_1$ we have

$$\beta_0 m_k \leq m_{k-1} \leq \beta_1 m_k.$$

Then there is a constant M independent of k such that

$$\eta_k^{-1} A_k(v, v) \leq A_k(B_k A_k v, v) \leq \eta_k A_k(v, v), \quad \forall v \in M_k$$

with

$$\eta_k = \frac{M + m_k^{\alpha}}{m_k^{\alpha}}.$$

Remark 6. *Condition* (3.10) *follows from Theorem 2 when $k = J+1$ and R_{J+1} is one of the smoothers from the theorem. For $k \leq J$, $M_k = \mathcal{V}_k^c$,* (3.10) *can be verified for the same smoothers in a way similar to the standard Galerkin case.*

From now on we will assume that $\Gamma_D = \partial\Omega$. We will use the following regularity assumption to prove (3.11): there exists $\rho \in (\frac{1}{2}, 1]$ and a constant C_Ω such that the solution, U, of the homogeneous problem (2.1) (i.e. when $u_D = 0$) satisfies

$$\|U\|_{1+\rho} \leq C_\Omega \|f\|_{-1+\rho}. \tag{3.12}$$

Lemma 5. *Assume that (3.12) holds. Then for all $u \in M_k$, $k = 2, \ldots, J+1$ we have*

$$|||u - P_{k-1}u|||_k \leq C h_k^\rho \|A_k u\|_{-1+\rho}.$$

(Here $|||\cdot|||_{J+1} = |||\cdot|||_J$.)

Proof. Let $w \in H^{1+\rho}(\Omega)$ be the solution of (2.1) with $f = A_k u$ and $u_D = 0$. Note that in this case

$$\mathcal{L}_\ell(v) = (A_k u, v), \quad \forall v \in L^2(\Omega).$$

We start by using the triangle inequality (with $|||\cdot|||_{*,J+1} = |||\cdot|||_{*,J}$)

$$|||u - P_{k-1}u|||_k \leq C\, |||u - P_{k-1}u|||_{*,k} \leq C\, |||u - w|||_{*,k} + C\, |||w - P_{k-1}u|||_{*,k}\,. \tag{3.13}$$

By the definition of A_k we have (with $\mathcal{A}_{J+1} = \mathcal{A}_J$ and $\mathcal{L}_{J+1} = \mathcal{L}_J$ if $k = J+1$)

$$\mathcal{A}_k(u, v) = (A_k u, v) = \mathcal{L}_k(v), \quad \forall v \in M_k.$$

Thus, u is the IP approximation of w from M_k and therefore by Lemma 4 we have

$$|||w - u|||_{*,k} \leq C h_k^\rho \|w\|_{1+\rho}. \tag{3.14}$$

By the definitions of P_{k-1} and A_k we have: $\forall v \in M_{k-1}$

$$\mathcal{A}_{k-1}(P_{k-1}u, v) = \mathcal{A}_k(u, v) = (A_k u, v) = \mathcal{L}_{k-1}(v)$$

which means that $P_{k-1}u$ is the IP approximation of w from M_{k-1} and so

$$|||w - P_{k-1}u|||_{*,k-1} \leq Ch_{k-1}^{\rho}\|w\|_{1+\rho}.$$

In this last estimate we want to replace $|||\cdot|||_{*,k-1}$ with $|||\cdot|||_{*,k}$ and h_{k-1} with h_k. We have two cases: 1) $k = J+1$ and 2) $k = 2, \ldots, J$. The first case is trivial since by definition

$$|||v|||_{*,J+1} = |||v|||_{*,J}\,, \quad \forall v \in H^{1+\rho}(\mathcal{T}_J), \quad \text{and} \quad h_{J+1} = h_J.$$

In the second case, we can write for $\ell = k-1, k$

$$|||v|||_{*,\ell}^2 \simeq (\nabla v, \nabla v) + \frac{\kappa}{h_\ell}\int_{\Gamma_D} v^2, \quad \forall v \in H^{1+\rho}(\mathcal{T}_\ell) \cap C(\overline{\Omega}).$$

Since $h_{k-1} \simeq h_k$ we have

$$|||v|||_{*,k-1} \simeq |||v|||_{*,k}\,, \quad \forall v \in H^{1+\rho}(\mathcal{T}_{k-1}) \cap C(\overline{\Omega}),$$

and therefore, since $(w - P_{k-1}u) \in H^{1+\rho}(\mathcal{T}_\ell) \cap C(\overline{\Omega})$,

$$|||w - P_{k-1}u|||_{*,k} \leq C\,|||w - P_{k-1}u|||_{*,k-1} \leq Ch_{k-1}^{\rho}\|w\|_{1+\rho} \leq Ch_k^{\rho}\|w\|_{1+\rho}. \tag{3.15}$$

Using (3.14) and (3.15) in (3.13) and then the regularity assumption (3.12) we obtain

$$|||u - P_{k-1}u|||_k \leq Ch_k^{\rho}\|w\|_{1+\rho} \leq Ch_k^{\rho}\|A_k u\|_{-1+\rho}$$

which completes the proof. □

Lemma 6. *Assume that (3.12) holds. Then condition (3.11) holds with $\alpha = \rho/2$.*

Proof. First we will show that

$$\|A_k u\|_{-1} \leq C\,|||u|||_k\,, \quad \forall u \in M_k. \tag{3.16}$$

Indeed,

$$\|A_k u\|_{-1} = \sup_{v\in H_0^1(\Omega)} \frac{(A_k u, v)}{|v|_1} \leq \sup_{v\in H_0^1(\Omega)} \frac{(A_k u, v-\Pi_k v)}{|v|_1} + \sup_{v\in H_0^1(\Omega)} \frac{(A_k u, \Pi_k v)}{|v|_1}.$$

Using the following two estimates for Π_k from [24]: for all $v \in H_0^1(\Omega)$

$$|\Pi_k v|_1 \leq C|v|_1$$

$$\|v - \Pi_k v\| \leq Ch_k|v|_1,$$

we get

$$\begin{aligned} \|A_k u\|_{-1} &\leq \sup_{v\in H_0^1(\Omega)} \frac{\|A_k u\|\,\|v-\Pi_k v\|}{|v|_1} + \sup_{v\in H_0^1(\Omega)} \frac{|||u|||_k\,|||\Pi_k v|||_k}{|v|_1} \\ &\leq Ch_k\|A_k u\| + C\,|||u|||_k\,, \end{aligned} \tag{3.17}$$

where for the second term we used the fact that $|||\Pi_k v|||_k \simeq |\Pi_k v|_1$. Since A_k is a symmetric and positive definite operator with respect to $(\cdot,\cdot)$ we have

$$\|A_k u\|^2 = (A_k^2 u, u) \leq \lambda_k(A_k u, u) \leq Ch_k^{-2}\,|||u|||_k^2\,.$$

Using this estimate in (3.17) we get (3.16).

The result of Lemma 5 gives

$$|||u - P_{k-1}u|||_k \leq Ch_k^{\rho}\|A_k u\|_{-1+\rho} \leq Ch_k^{\rho}\|A_k u\|_{-1}^{1-\rho}\|A_k u\|^{\rho}. \tag{3.18}$$

For the last estimate we used the fact that $H^{-1+\rho}(\Omega)$ is an intermediate space between

$H^{-1}(\Omega)$ and $L^2(\Omega)$. Now, using (3.18) and (3.16) we obtain

$$\begin{aligned}
|A_k(u - P_{k-1}u, u)| &\leq |||u - P_{k-1}u|||_k \, |||u|||_k \\
&\leq C h_k^{\rho} \|A_k u\|_{-1}^{1-\rho} \|A_k u\|^{\rho} \, |||u|||_k \\
&\leq C h_k^{\rho} \, |||u|||_k^{1-\rho} \, \|A_k u\|^{\rho} \, |||u|||_k \\
&\leq C \frac{\|A_k u\|^{\rho}}{\lambda_k^{\rho/2}} \, |||u|||_k^{2-\rho} \\
&= C \left(\frac{\|A_k u\|^2}{\lambda_k} \right)^{\frac{\rho}{2}} A_k(u,u)^{1-\frac{\rho}{2}}.
\end{aligned}$$

This is exactly (3.11) with $\alpha = \rho/2$. □

3.2.3. Numerical Experiments

We consider the same three Test Problems we used for the numerical experiments for the two-level preconditioners (see page 32 for the description of the tests). All test runs we present here are multilevel extensions of their corresponding two-level experiments and therefore we can compare the two in order to measure the quality of the multigrid algorithms. The test setup is the same as before: we use the same iterative method (PCG), stopping criterion, smoother, values for κ. Once again, we generate a sequence of nested meshes starting with a coarse ("Level 0") mesh and using k times uniform refinement to obtain the "Level k" mesh. We use this mesh hierarchy to define the multigrid algorithms as described earlier in the chapter (see page 40, note that here we start counting the levels with "0", not "1"). The coarsest ("Level 0") problems are solved using the PCG method with Gauss-Seidel preconditioner and the same relative accuracy. As before, we will report the number of iterations ("iter") and the average reduction factors ("arf"). To completely, define the multigrid preconditioners we need to specify the number of correction and smoothing steps. We use the following three choices:

- V-cycle: use $p = 1$ correction and $m_k = 1$ pre- and post-smoothing steps at all levels $k = 1, \ldots, J$.
- variable V-cycle: use $p = 1$ correction and $m_k = 2^{J-k}$ pre- and post-smoothing steps, $k = 1, \ldots, J$. Here J is the level at which we solve the discrete problem.
- W-cycle: use $p = 2$ correction and $m_k = 1$ pre- and post-smoothing steps, $k = 1, \ldots, J$.

The numerical results for Test Problem 1 are presented in Tables 3.5 and 3.6 for linear and quadratic finite elements, respectively. The rows labeled "$\mathcal{S}$ dof" give the number of degrees of freedom for the space $\mathcal{S}$ which is one of $\mathcal{V}_J$, $\mathcal{V}_J^c$, or $\overline{\mathcal{V}}_J$. The rows below the "$\mathcal{S}$ dof" row give results for the method based on the space $\mathcal{S}$ using different *-cycle methods. The results are in the form "iter/arf".

Table 3.5. Multigrid preconditioners, Test Problem 1, linear FE

	Level 2	Level 3	Level 4	Level 5	Level 6
$\mathcal{V}_J$ dof	3,072	24,576	196,608	1,572,864	12,582,912
$\mathcal{V}_J^c$ dof	189	1,241	9,009	68,705	536,769
V-cycle	13/0.2322	13/0.2365	13/0.2337	13/0.2252	12/0.2124
var. V-cycle	13/0.2309	13/0.2338	13/0.2274	12/0.2134	12/0.2013
$\overline{\mathcal{V}}_J$ dof	768	6,144	49,152	393,216	3,145,728
V-cycle	21/0.4022	30/0.5333	40/0.6229	51/0.6915	62/0.7426
var. V-cycle	20/0.3897	26/0.4795	29/0.5188	30/0.5325	31/0.5434
W-cycle	20/0.3845	24/0.4535	24/0.4632	24/0.4623	24/0.4588

For Method I we test the V-cycle and variable V-cycle methods. For both linear and quadratic elements, both *-cycle methods give almost identical number of

Table 3.6. Multigrid preconditioners, Test Problem 1, quadratic FE

	Level 1	Level 2	Level 3	Level 4	Level 5
$\mathcal{V}_J$ dof	960	7,680	61,440	491,520	3,932,160
$\mathcal{V}_J^c$ dof	189	1,241	9,009	68,705	536,769
V-cycle	10/0.1414	11/0.1729	11/0.1725	11/0.1655	10/0.1547
var. V-cycle	10/0.1329	10/0.1538	10/0.1469	10/0.1351	9/0.1179
$\overline{\mathcal{V}}_J$ dof	96	768	6,144	49,152	393,216
V-cycle	19/0.3677	29/0.5273	41/0.6282	52/0.6963	62/0.7416
var. V-cycle	19/0.3616	28/0.5062	36/0.5867	38/0.6143	39/0.6207
W-cycle	19/0.3623	28/0.5108	35/0.5792	35/0.5890	35/0.5880

iterations independent of the refinement level. Comparing with the results for the two-level Method I (Tables 3.1 and 3.2) we can see that the multilevel methods are extremely close to the two-level one since the number of iterations increases by at most 2. Note that in our theoretical analysis we proved that the variable V-cycle Method I gives rise to a uniform preconditioner but we do not have a proof for the V-cycle Method I, even though our tests indicate that.

For Method II, for which we do not have any theoretical results, we test all three V-, variable V-, and W-cycle methods. The results are presented in the bottom halves of Tables 3.5 and 3.6. We can clearly see that for the V-cycle Method II the number of iterations increases linearly with the refinement level. In the other two cases we observe that the number of iterations stabilizes as we refine the mesh. In the case of W-cycle this stabilization occurs at a coarser level compared to the variable V-cycle and the former always uses less iterations to converge than the latter. A comparison with the two-level Method II (Tables 3.1 and 3.2) shows that the W-cycle Method II is

very close to the two-level preconditioner in that it uses at most 20% more iterations. Note that all these observations are valid for both linear and quadratic elements.

Comparing the results for Method I with those for W-cycle Method II we see that the former uses about 2 times less iterations for linear and about 3.5 times less iterations for quadratic elements than the latter. This clearly shows the advantage (V- or variable V-cycle) Method I has over W-cycle Method II for Test Problem 1.

For Test Problem 2 we test only the V-cycle Method I and the W-cycle Method II. The numerical results are presented in Table 3.7. From the results we see that both preconditioners perform very similarly to their corresponding two-level counterparts (see Table 3.3): we observe a small increase in the number of iterations in all cases. The W-cycle Method II still converges in less iterations than V-cycle Method I for small values of ϵ, even though the gap between the two is smaller than it is for the corresponding two-level methods. To summarize, both multilevel methods converge uniformly in both h and ϵ with the W-cycle Method II having a small advantage over the V-cycle Method I for small values of ϵ.

The numerical results for Test Problem 3 are presented in Table 3.8. Once again we test the V-cycle Method I and the W-cycle Method II. If we compare these results with the corresponding two-level methods we see that there is almost no increase in the number of iterations for Method I and only a small increase for Method II. Due to the geometrical anisotropies introduced by the discretization of the domain we can observe that the number of iterations are around 2 times larger compared to those we see for the regular mesh used in Test Problem 1 (cf. Table 3.5). Clearly, the V-cycle Method I is much better than the W-cycle Method II for this Test Problem since the former needs less than half the iterations that the latter needs to converge.

Table 3.7. Multigrid preconditioners, Test Problem 2, linear FE

	Level 1	Level 2	Level 3	Level 4	Level 5
$\mathcal{V}_J$ dof	1,344	10,752	86,016	688,128	5,505,024
$\mathcal{V}_J^c$ dof	117	665	4,401	31,841	241,857
$\epsilon = 1$	14/0.2443	14/0.2657	14/0.2634	14/0.2560	14/0.2480
$\epsilon = 0.1$	15/0.2827	18/0.3512	20/0.3844	20/0.3870	20/0.3906
$\epsilon = 0.01$	17/0.3377	21/0.4047	26/0.4796	28/0.5061	29/0.5286
$\epsilon = 0.001$	17/0.3329	21/0.4094	26/0.4883	29/0.5164	31/0.5480
$\epsilon = 10^{-4}$	17/0.3202	21/0.4058	26/0.4825	28/0.5165	29/0.5236
$\epsilon = 10^{-5}$	16/0.3014	21/0.4020	24/0.4609	27/0.4943	28/0.5153
$\epsilon = 10^{-6}$	15/0.2911	20/0.3832	23/0.4407	25/0.4758	26/0.4913
$\overline{\mathcal{V}}_J$ dof	336	2,688	21,504	172,032	1,376,256
$\epsilon = 1$	18/0.3527	22/0.4307	24/0.4544	24/0.4561	24/0.4518
$\epsilon = 0.1$	20/0.3791	23/0.4302	24/0.4577	24/0.4626	24/0.4626
$\epsilon = 0.01$	19/0.3721	22/0.4316	24/0.4578	25/0.4695	25/0.4745
$\epsilon = 0.001$	19/0.3710	23/0.4352	24/0.4592	25/0.4712	25/0.4768
$\epsilon = 10^{-4}$	19/0.3676	22/0.4300	24/0.4594	25/0.4715	25/0.4771
$\epsilon = 10^{-5}$	19/0.3644	21/0.4095	24/0.4594	25/0.4715	25/0.4771
$\epsilon = 10^{-6}$	18/0.3542	21/0.4033	24/0.4594	25/0.4715	25/0.4771

Table 3.8. Multigrid preconditioners, Test Problem 3, linear FE

	Level 1	Level 2	Level 3	Level 4
$\mathcal{V}_J$ dof	24,032	192,256	1,538,048	12,304,384
$\mathcal{V}_J^c$ dof	1,445	9,693	70,633	538,513
iter/arf	18/0.3530	18/0.3559	18/0.3529	19/0.3785
$\overline{\mathcal{V}}_J$ dof	6,008	48,064	384,512	3,076,096
iter/arf	35/0.5907	40/0.6307	45/0.6578	48/0.6788

CHAPTER IV

ALGEBRAIC MULTIGRID METHODS

In this chapter we introduce and study numerically an algebraic multigrid (AMG) algorithm for the preconditioning of the SIPG method. In addition to the assembled matrix of the SIPG method, our AMG method only requires basic topological information about the mesh and therefore can be used on arbitrary unstructured meshes. The basic idea is to define coarse spaces of piecewise constant functions in order to simulate a space hierarchy similar to the one used by multigrid Method II defined in the previous chapter which in the case of uniform refinement is readily available. We build a sequence of nested coarse "triangulations" based on an element agglomeration algorithm that uses only the topology of the initial finest mesh. The corresponding piecewise constant spaces are nested and we use the natural embeddings to construct the inherited coarse level matrices. We also consider a smoothed aggregation/interpolation version of the algorithm combined with more aggressive coarsening. This approach leads to improved convergence rates. In the numerical experiments we study not only the convergence properties of the preconditioners but also the computational complexity of their construction and usage.

4.1. Element Agglomeration AMG

We will use the following algorithm to construct our AMG preconditioner: assume that we are given the finest level matrix A_J of size $(n_J \times n_J)$ and the prolongation matrices P_k of sizes $(n_k \times n_{k-1})$, for $k = 1, \ldots, J$; here n_k denotes the dimension of the k-th level. We define the coarse matrices recursively using an RAP matrix multiplication:

$$A_{k-1} = P_k^t A_k P_k, \qquad k = J, \ldots, 1.$$

Now, given the matrix A_k we can define a smoothing matrix R_k, for example $R_k = (D_k + L_k^t)^{-1} D_k (D_k + L_k)^{-1}$ is the symmetric Gauss-Seidel smoother; here D_k is the diagonal of A_k, and L_k is the strictly lower triangular part of A_k. In general R_k need not be symmetric and therefore we denote

$$R_k^{(\ell)} = \begin{cases} R_k & \text{if } \ell \text{ is odd,} \\ R_k^t & \text{if } \ell \text{ is even.} \end{cases}$$

Given a number m_k of pre- and post-smoothing steps and a number p of correction steps we define the multigrid preconditioner B_J recursively: set $B_0 = A_0^{-1}$; then for $k = 1, \ldots, J$ define the action of B_k on a given vector $g \in \mathbb{R}^{n_k}$ by

1. pre-smoothing: define $x^\ell \in \mathbb{R}^{n_k}$, $\ell = 0, \ldots, m_k$ by: set $x^0 = 0$ and

$$x^\ell = x^{\ell-1} + R_k^{(\ell+m_k)}(g - A_k x^{\ell-1}), \quad \ell = 1, \ldots, m_k$$

2. correction: define $y^{m_k} = x^{m_k} + P_k q^p$ where $q^\ell \in \mathbb{R}^{n_{k-1}}$, $\ell = 0, \ldots, p$ are defined by: set $q^0 = 0$, and

$$q^\ell = q^{\ell-1} + B_{k-1}[P_k^t(g - A_k x^{m_k}) - A_{k-1} q^{\ell-1}], \quad \ell = 1, \ldots, p.$$

3. post-smoothing: define $y^\ell \in \mathbb{R}^{n_k}$, $\ell = m_k + 1, \ldots, 2m_k$ by

$$y^\ell = y^{\ell-1} + R_k^{(\ell+m_k)}(g - A_k y^{\ell-1}).$$

4. $B_k g = y^{2m_k}$.

Remark 7. *In the multilevel setup from the previous chapter we can define the prolongation matrices P_k from the embeddings $\overline{\mathcal{V}}_1 \subset \cdots \subset \overline{\mathcal{V}}_J \subset \mathcal{V}_J$. In this case the AMG algorithm we just described differs from Method II in the previous chapter only*

in the choice of the coarse bilinear forms. Here we use the bilinear form $\mathcal{A}_J(\cdot,\cdot)$ on all levels which is the nested-inherited case.

In order to construct the prolongation (interpolation) matrices P_k we will define a sequence of nested triangulations $\{\mathcal{T}_k\}_{k=1}^J$ and use the natural embeddings of the corresponding piecewise constant spaces $\left\{\overline{\mathcal{V}}_k\right\}_{k=1}^J$. The finest mesh $\mathcal{T}_J$ is the mesh on which the matrix A_J is assembled on the discontinuous space $\mathcal{V}_J$ of piecewise polynomial functions. We define the matrix P_J to be the matrix representation of the embedding $\overline{\mathcal{V}}_J \subset \mathcal{V}_J$ in the standard bases for both spaces. For example, for linear elements P_J has the block form

$$P_J = \begin{pmatrix} e & & 0 \\ & \ddots & \\ 0 & & e \end{pmatrix} \quad \text{where} \quad e = (1,1,1,1)^t.$$

To construct the mesh hierarchy $\{\mathcal{T}_k\}_{k=1}^J$ we assume that we have an enumeration of the elements in the finest mesh $\mathcal{T}_J = \{T_{J,1}, T_{J,2}, \ldots, T_{J,m}\}$ where $m = n_{J-1}$ and the following $(m \times m)$ adjacency matrix:

$$H_{ij} = \begin{cases} 1 & \text{if } i \neq j \text{ and } T_{J,i},\ T_{J,j} \text{ are neighbors,} \\ 0 & \text{otherwise.} \end{cases}$$

We consider two elements to be neighbors if they have a common face. Note that H has the sparsity pattern of A_{J-1} excluding the diagonal. Given the matrix H we define the auxiliary mesh hierarchy $\{\mathcal{T}_k^*\}$ where $\mathcal{T}_k^*$ has exactly 2^k elements. The elements in $\mathcal{T}_k^*$ are defined by a partition of the elements of the mesh $\mathcal{T}_J$ — every element $T_{J,i}$ belongs to exactly one element of $\mathcal{T}_k^*$. The mesh $\mathcal{T}_0^*$ has 1 element which is the union of all elements in $\mathcal{T}_J$. Given the mesh $\mathcal{T}_k^*$ we define the next mesh $\mathcal{T}_{k+1}^*$ by splitting every element $T \in \mathcal{T}_k^*$ into two elements: let $\ell_1, \ell_2, \ldots, \ell_s$ be the indexes

of the elements in $\mathcal{T}_J$ whose union is T:

$$T = \bigcup_{i=1}^{s} T_{J,\ell_i}$$

and define the $(s \times s)$ adjacency matrix of T:

$$H_{ij}^{(T)} = H_{\ell_i \ell_j}.$$

We use a graph bisection algorithm to split the graph defined by $H^{(T)}$. The result, in the form of a binary vector $b \in \{0,1\}^s$, defines the two new elements $T_0, T_1 \in \mathcal{T}_{k+1}^*$:

$$T_i = \bigcup_{j:b_j=i} T_{J,\ell_j}, \qquad i = 0, 1.$$

We use the graph partitioning library METIS (routine `METIS_PartGraphRecursive`) as our bisection algorithm. Sometimes the elements T_0 and/or T_1 produced by METIS are not connected which we want to avoid and therefore in those rare cases we use a simpler bisection algorithm that generates connected elements. The process of generating the auxiliary meshes $\mathcal{T}_k^*$ is terminated when $2^k \geq m/\theta$ where $\theta = 2^\alpha$ is a given coarsening factor. Let ℓ be the smallest integer such that $2^\ell \geq m/\theta$ then $J = [\ell/\alpha] + 2$ and we define

$$\mathcal{T}_{J-1} = \mathcal{T}_\ell^*, \quad \mathcal{T}_{J-2} = \mathcal{T}_{\ell-\alpha}^*, \quad \ldots \quad \mathcal{T}_1 = \mathcal{T}_{\ell-[\ell/\alpha]\alpha}^*.$$

For example, if $m = 1000$, $\theta = 8$ ($\alpha = 3$) then $\ell = 7$, $J = 4$ and $\mathcal{T}_3 = \mathcal{T}_7^*$, $\mathcal{T}_2 = \mathcal{T}_4^*$, $\mathcal{T}_1 = \mathcal{T}_1^*$; the level dimensions are $n_4 = 4000$ (assuming linear elements), $n_3 = 1000$, $n_2 = 128$, $n_1 = 16$, $n_0 = 2$.

To illustrate the algorithm, on Figure 4.1 we show the first two auxiliary meshes $\mathcal{T}_1^*$ and $\mathcal{T}_2^*$ obtained when we apply our algorithm to the mesh of Test Problem 3 (see page 32). The different colors represent the different agglomerated elements.

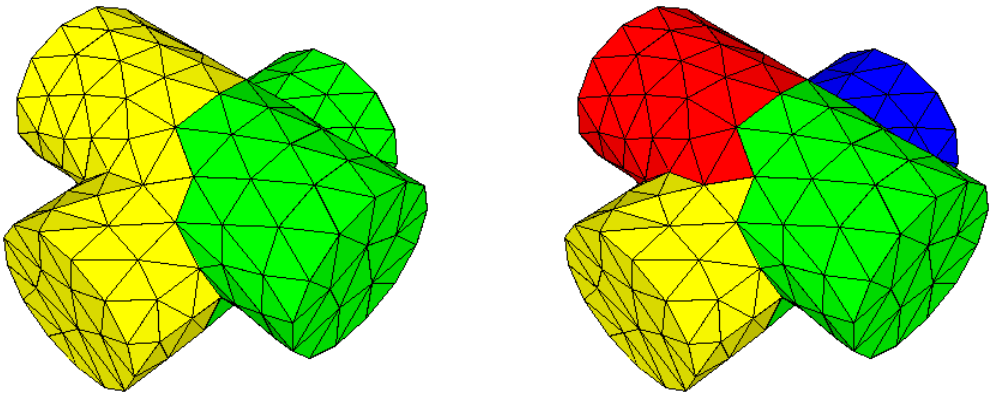

Fig. 4.1. Auxiliary agglomerated triangulations: $\mathcal{T}_1^*$ (left) and $\mathcal{T}_2^*$ (right).

Remark 8. *Note that even if the bisection algorithm is optimal, the complexity of the algorithm to construct the hierarchy* $\{\mathcal{T}_k\}_{k=1}^J$ *is at least* $\mathcal{O}(m \log m)$.

Having defined the sequence of nested meshes $\{\mathcal{T}_k\}_{k=1}^J$, we can define their corresponding piecewise constant spaces $\left\{\overline{\mathcal{V}}_k\right\}_{k=1}^J$. Then the prolongation matrix P_k for $k = 1, \ldots, J-1$ is defined to be the representation of the embedding $\overline{\mathcal{V}}_k \subset \overline{\mathcal{V}}_{k+1}$.

4.2. Smoothed Aggregation

In this section we consider a modification of the AMG preconditioner described above that is aimed at improving its convergence properties. The approach we present here was first introduced in [25] and later analyzed in [26]. The idea is to smooth the prolongation matrices $\{P_k\}$ using the corresponding coarse matrices. Thus, we replace the general prolongation P_k which was constructed using only the topology of the mesh $\mathcal{T}_J$ with an improved prolongation $\widetilde{P}_k$ which is designed specifically for the fine level matrix A_J. We define $\widetilde{P}_k$ and the corresponding coarse matrices $\widetilde{A}_k$ as follows: set $\widetilde{A}_J = A_J$; due to the large number of nonzero entries per row in A_J we

choose to not smooth the prolongation P_J:

$$\widetilde{P}_J = P_J \qquad \widetilde{A}_{J-1} = A_{J-1} = P_J^t A_J P_J.$$

The rest of the prolongations and matrices we define recursively:

$$\widetilde{P}_k = (I - \lambda_k^{-1}\widetilde{A}_k)P_k \qquad \widetilde{A}_{k-1} = \widetilde{P}_k^t \widetilde{A}_k \widetilde{P}_k, \qquad k = J-1, \ldots, 1$$

where λ_k is chosen so that $\rho(I - \lambda_k^{-1}\widetilde{A}_k) < 1$. In the numerical experiments we used

$$\lambda_k = \frac{1}{2}\max_i \sum_j |(\widetilde{A}_k)_{ij}| = \frac{1}{2}\|\widetilde{A}_k\|_\infty = \frac{1}{2}\|\widetilde{A}_k\|_1 \geq \frac{1}{2}\lambda_{\max}(\widetilde{A}_k).$$

This smoothing method can be viewed as a process in which we replace the piecewise constant spaces $\overline{\mathcal{V}}_k$ with a space spanned by the basis functions given by the columns of the matrix product $\widetilde{P}_J \widetilde{P}_{J-1} \cdots \widetilde{P}_k$. Each such basis function is associated with an element in the triangulation $\mathcal{T}_k$, however its support is larger than that element. This increase in the support of the basis functions leads to an increase in the sparsity pattern of the matrices $\widetilde{A}_k$. In order to control the sparsity one can use more aggressive coarsening: $\theta = 16$, $\theta = 32$ instead of the standard (for 3D) $\theta = 8$. Finally, we define the smoothed aggregation AMG preconditioner $\widetilde{B}_J$ in the same way we defined B_J replacing P_k and A_k with $\widetilde{P}_k$ and $\widetilde{A}_k$, respectively.

4.3. Numerical Experiments

We use the same test setup as in the previous chapter, see pages 32, 50. In the numerical experiments presented here we study the properties of the AMG preconditioners on a sequence of geometrically refined meshes. Note that this mesh hierarchy is not used in the construction of the preconditioners, instead we use the agglomeration algorithm described above. We will use M1 to refer to the AMG preconditioner B_J

Table 4.1. AMG preconditioners, Test Problem 1, linear FE

	Level 2	Level 3	Level 4	Level 5	Level 6
M1, $\theta = 8$, v. V-cycle	20/0.3893	26/0.4827	32/0.5561	41/0.6357	51/0.6944
M1, $\theta = 8$, W-cycle	20/0.3803	24/0.4490	26/0.4917	30/0.5318	33/0.5625
M2, $\theta = 8$, V-cycle	20/0.3788	22/0.4303	24/0.4537	25/0.4677	25/0.4745
M2, $\theta = 16$, V-cycle	20/0.3785	24/0.4526	26/0.4867	30/0.5341	31/0.5443
M2, $\theta = 32$, V-cycle	21/0.4066	27/0.4950	31/0.5443	37/0.6023	42/0.6398
M2, $\theta = 8$, W-cycle	20/0.3717	21/0.4089	21/0.4056	21/0.4045	21/0.4030
M2, $\theta = 16$, W-cycle	20/0.3718	22/0.4173	22/0.4191	22/0.4256	22/0.4261
M2, $\theta = 32$, W-cycle	20/0.3801	24/0.4480	24/0.4629	26/0.4796	26/0.4873

based on the prolongation matrices P_k, and M2 — to the preconditioner $\widetilde{B}_J$ (based on $\widetilde{P}_k$).

We first consider Test Problem 1 discretized with linear finite elements. In Table 4.1 we present the number of iterations and average reduction factors from the PCG method for both AMG preconditoners (M1 and M2) using the indicated coarsening factor θ and cycle (V, variable V, or W). We observe that for M1 in both variable V- and W-cycle the number of iterations grows with the refinement level, however the increase in the numbers is much slower for the W-cycle. When smoothed aggregation is used (M2) the number of iterations remains bounded in all cases except possibly the case $\theta = 32$ using V-cycle. As we can expect, the W-cycle gives better convergence rates than the V-cycle and using more aggressive coarsening (larger θ) slows the convergence.

We will use the following ratios to express the computational cost of one V-cycle

Table 4.2. Complexity of AMG preconditioners, Test Problem 1, linear FE

	Level 2	Level 3	Level 4	Level 5	Level 6
M1, $\theta = 8$	1.079/1.192	1.081/1.210	1.082/1.216	1.083/1.221	1.083/1.224
M2, $\theta = 8$	1.103/1.301	1.121/1.426	1.136/1.557	1.149/1.726	1.160/1.919
M2, $\theta = 16$	1.079/1.186	1.084/1.215	1.088/1.236	1.090/1.254	1.092/1.267
M2, $\theta = 32$	1.071/1.154	1.073/1.164	1.075/1.172	1.076/1.177	1.076/1.180

relative to the cost of a matrix-vector product with the fine level matrix A_J:

$$\varkappa_v = \frac{1}{\eta(A_J)} \sum_{k=0}^{J} \eta(A_k),$$

where $\eta(A)$ denotes the number of nonzero entries in the sparse matrix A. For W-cycle we use the ratio:

$$\varkappa_w = \frac{1}{\eta(A_J)} \sum_{k=0}^{J} 2^{J-k} \eta(A_k).$$

In Table 4.2 we give these relative complexities (in the form "$\varkappa_v/\varkappa_w$") for Test Problem 1 using linear elements. In all cases the complexities are small and do not increase substantially with the refinement level. The only exception is when smoothed aggregation is used (M2) with coarsening factor $\theta = 8$ and even in this case the increase is noticeable only for W-cycle. This effect was expected and it is due to the increased number of nonzero entries per row in the coarse matrices $\widetilde{A}_k$. As we expected, using aggressive coarsening ($\theta = 16, 32$) resolved this problem.

The algorithm for constructing the AMG preconditioners can be divided into two steps: 1) element agglomeration or construction of the prolongation matrices P_k and 2) construction of the coarse matrices A_k or $\widetilde{A}_k$ (including the construction of $\widetilde{P}_k$). As we noted in Remark 8 our algorithm for step 1 is not optimal, however in our tests we observed that it is not much slower than step 2 even for the largest problems. To

Table 4.3. Setup cost of AMG preconditioners, Test Problem 1, linear FE

	Level 2	Level 3	Level 4	Level 5	Level 6
M1, $\theta = 8$	1.376	1.382	1.384	1.385	1.386
M2, $\theta = 8$	1.972	2.557	3.440	5.032	7.559
M2, $\theta = 16$	1.580	1.712	1.809	1.910	1.983
M2, $\theta = 32$	1.501	1.562	1.597	1.623	1.638

measure the complexity of step 2 we compute the complexity of all matrix-matrix products involved in the step and divide by $\eta(A_J)$: for M1 we use

$$\varkappa_s = \frac{1}{\eta(A_J)} \sum_{k=1}^{J} \left[\mu(A_k, P_k) + \mu(P_k^t, A_k P_k) \right],$$

where $\mu(A, B)$ denotes the number of floating point multiplications needed to perform the matrix-matrix multiplication of the sparse matrices A and B; for M2 we compute

$$\varkappa_s = \frac{1}{\eta(A_J)} \sum_{k=1}^{J} \left[\mu(S_k, P_k) + \mu(\widetilde{A}_k, \widetilde{P}_k) + \mu(\widetilde{P}_k^t, \widetilde{A}_k \widetilde{P}_k) \right],$$

where $S_k = I - \lambda_k^{-1} \widetilde{A}_k$ is the prolongation smoother ($\widetilde{P}_k = S_k P_k$). The results are presented in Table 4.3. As we can expect due to the increasing sparsity pattern of the coarse matrices $\widetilde{A}_k$, the relative setup complexity of the smoothed aggregation AMG (M2) preconditioner with standard coarsening factor $\theta = 8$ increases substantially with the refinement level. As we see from the results, using aggressive coarsening reduces the setup cost significantly. In all cases, except [M2, $\theta = 8$], the relative setup complexities increase slowly but remain relatively small and bounded.

We next consider Test Problem 3 discretized with linear finite elements. In Table 4.4 we present the number of PCG iterations and average reduction factors for the indicated AMG preconditioners. In all cases we observe a linear increase in the

Table 4.4. AMG preconditioners, Test Problem 3, linear FE

	Level 1	Level 2	Level 3	Level 4
M1, $\theta = 8$, W-cycle	35/0.5903	43/0.6465	49/0.6860	56/0.7163
M2, $\theta = 8$, V-cycle	36/0.5897	41/0.6370	46/0.6687	51/0.6940
M2, $\theta = 16$, V-cycle	37/0.6019	45/0.6598	52/0.6974	60/0.7331
M2, $\theta = 32$, V-cycle	39/0.6147	49/0.6851	60/0.7341	72/0.7713
M2, $\theta = 8$, W-cycle	34/0.5786	38/0.6125	41/0.6348	44/0.6547
M2, $\theta = 16$, W-cycle	35/0.5835	39/0.6229	43/0.6481	46/0.6682
M2, $\theta = 32$, W-cycle	36/0.5906	42/0.6413	47/0.6731	51/0.6940

Table 4.5. Complexity of AMG preconditioners, Test Problem 3, linear FE

	Level 1	Level 2	Level 3	Level 4
M1, $\theta = 8$	1.084/1.219	1.084/1.226	1.084/1.228	1.084/1.229
M2, $\theta = 8$	1.125/1.440	1.140/1.579	1.151/1.741	1.161/1.924
M2, $\theta = 16$	1.085/1.219	1.089/1.242	1.090/1.255	1.091/1.266
M2, $\theta = 32$	1.073/1.163	1.074/1.170	1.075/1.174	1.075/1.176

iterations with the refinement level. This is not unexpected because similar behavior can be observed for the two-level method using the coarse space $\overline{\mathcal{V}}$ (see the last row in Table 3.4 on page 37). If we compare the results with those for Test Problem 1 we see that the convergence for this Test Problem is slower which can be explained with the geometrical anisotropies of the mesh. In Tables 4.5 and 4.6 we give the relative complexities to apply the preconditioner and for the setup stage, respectively. The results are very similar to those for Test Problem 1 — the complexities are small and bounded for all cases except the smoothed aggregation AMG (M2) preconditioner

Table 4.6. Setup cost of AMG preconditioners, Test Problem 3, linear FE

	Level 1	Level 2	Level 3	Level 4
M1, $\theta = 8$	1.384	1.386	1.386	1.386
M2, $\theta = 8$	2.589	3.594	5.143	7.726
M2, $\theta = 16$	1.747	1.860	1.940	2.005
M2, $\theta = 32$	1.572	1.608	1.629	1.643

using coarsening factor $\theta = 8$.

CHAPTER V

THE METHOD OF BAUMANN AND ODEN

For functions u and v in $H^s(\mathcal{T})$, $s > \frac{3}{2}$, the bilinear form of the method of Baumann and Oden is

$$\mathcal{A}(u,v) = (a\nabla u, \nabla v) + \langle \{a\nabla v \cdot \mathbf{n}\}, [\![u]\!] \rangle_{\mathcal{E}_i \cup \mathcal{E}_D} - \langle \{a\nabla u \cdot \mathbf{n}\}, [\![v]\!] \rangle_{\mathcal{E}_i \cup \mathcal{E}_D}$$

and the its linear form is

$$\mathcal{L}(v) = (f,v) + \langle u_N, v \rangle_{\mathcal{E}_N} + \langle u_D, a\nabla v \cdot \mathbf{n} \rangle_{\mathcal{E}_D}.$$

With these definitions, the Baumann-Oden discontinuous Galerkin method for our elliptic problem (2.1) reads: find $u \in \mathcal{V}$ such that

$$\mathcal{A}(u,v) = \mathcal{L}(v), \quad \forall v \in \mathcal{V}. \tag{5.1}$$

5.1. Mixed Formulation

We consider the following L^2-orthogonal decomposition of the discrete space $\mathcal{V}$ into a direct sum

$$\mathcal{V} = \overline{\mathcal{V}} \oplus \mathcal{V}_0$$

where

$$\overline{\mathcal{V}} = \left\{ v \in L^2(\Omega) : v|_T = \text{const}, \ \forall T \in \mathcal{T} \right\}$$
$$\mathcal{V}_0 = \left\{ v \in \mathcal{V} : (v,w) = 0, \ \forall w \in \overline{\mathcal{V}} \right\}$$

that is $\overline{\mathcal{V}}$ is the space of piecewise constant functions and $\mathcal{V}_0$ is the space of the piecewise polynomial functions (of degree r) with average 0 over each element. Consider

the restriction of the form $\mathcal{A}(\cdot,\cdot)$ to $\mathcal{V}_0 \times \overline{\mathcal{V}}$ which we will denote by $\mathcal{B}(\cdot,\cdot)$

$$\mathcal{B}(v_0,\bar{u}) = \mathcal{A}(\bar{u},v_0) = \langle \{a\nabla v_0 \cdot \mathbf{n}\}, [\![\bar{u}]\!] \rangle_{\mathcal{E}_i \cup \mathcal{E}_D}, \qquad \forall \bar{u} \in \overline{\mathcal{V}},\ \forall v_0 \in \mathcal{V}_0.$$

Note that $\forall \bar{v} \in \overline{\mathcal{V}}$ and $\forall u_0 \in \mathcal{V}_0$ we have

$$\mathcal{A}(u_0,\bar{v}) = -\langle \{a\nabla u_0 \cdot \mathbf{n}\}, [\![\bar{v}]\!] \rangle_{\mathcal{E}_i \cup \mathcal{E}_D} = -\mathcal{A}(\bar{v},u_0) = -\mathcal{B}(u_0,\bar{v})$$

and also $\forall \bar{u}, \bar{v} \in \overline{\mathcal{V}}$ we have

$$\mathcal{A}(\bar{u},\bar{v}) = 0.$$

Thus, if we write $u = u_0 + \bar{u}$ and $v = v_0 + \bar{v}$ with $u_0, v_0 \in \mathcal{V}_0$ and $\bar{u}, \bar{v} \in \overline{\mathcal{V}}$ then we have

$$\mathcal{A}(u,v) = \mathcal{A}(u_0,v_0) + \mathcal{B}(v_0,\bar{u}) - \mathcal{B}(u_0,\bar{v}).$$

Therefore, the discrete problem (5.1) can be written in the following equivalent mixed form: find $u_0 \in \mathcal{V}_0$ and $\bar{u} \in \overline{\mathcal{V}}$ such that

$$\begin{aligned} \mathcal{A}(u_0,v_0) + \mathcal{B}(v_0,\bar{u}) &= \mathcal{L}(v_0), \quad \forall v_0 \in \mathcal{V}_0 \\ -\mathcal{B}(u_0,\bar{v}) &= \mathcal{L}(\bar{v}), \quad \forall \bar{v} \in \overline{\mathcal{V}}. \end{aligned} \tag{5.2}$$

We will use the following inner product and corresponding norm on $H^1(\mathcal{T})$

$$\begin{aligned} \langle\!\langle u,v \rangle\!\rangle &= (a\nabla u, \nabla v) + \left\langle \kappa h_{\mathcal{E}}^{-1} a_{\mathcal{E}} [\![\bar{u}]\!], [\![\bar{v}]\!] \right\rangle_{\mathcal{E}_i \cup \mathcal{E}_D}, && \forall u,v \in H^1(\mathcal{T}), \\ |||u||| &= \langle\!\langle u,u \rangle\!\rangle^{\frac{1}{2}}, && \forall u \in H^1(\mathcal{T}), \end{aligned}$$

where $\bar{u}$ is the L^2-orthogonal projection of u onto $\overline{\mathcal{V}}$ (that is the element-by-element average of u). Here the parameter $\kappa > 0$ is arbitrary. Note that the spaces $\mathcal{V}_0$ and $\overline{\mathcal{V}}$ are orthogonal with respect to the inner product $\langle\!\langle \cdot,\cdot \rangle\!\rangle$. When restricted to $\mathcal{V}_0$ and $\overline{\mathcal{V}}$

the norm simplifies to

$$\begin{aligned} |||u_0|||^2 &= (a\nabla u_0, \nabla u_0), && \forall u_0 \in \mathcal{V}_0 \\ |||\bar{u}|||^2 &= \left\langle \kappa h_{\mathcal{E}}^{-1} a_{\mathcal{E}} \left[\!\left[\bar{u}\right]\!\right], \left[\!\left[\bar{u}\right]\!\right] \right\rangle_{\mathcal{E}_i \cup \mathcal{E}_D}, && \forall \bar{u} \in \overline{\mathcal{V}}. \end{aligned}$$

We will study the mixed problem (5.2) using the following theorem (this is a version of Theorem 1.1 on page 42 in [11])

Theorem 6. *Assume that the bilinear forms $\mathcal{A}(\cdot,\cdot)$ and $\mathcal{B}(\cdot,\cdot)$ are bounded:*

$$\begin{aligned} \mathcal{A}(u_0, v_0) &\leq \alpha_1 |||u_0||| \, |||v_0|||, && \forall u_0, v_0 \in \mathcal{V}_0 \\ \mathcal{B}(u_0, \bar{v}) &\leq \beta_1 |||u_0||| \, |||\bar{v}|||, && \forall u_0 \in \mathcal{V}_0, \forall \bar{v} \in \overline{\mathcal{V}}, \end{aligned}$$

the bilinear form $\mathcal{A}(\cdot,\cdot)$ is coercive on $\mathcal{V}_0$: there exists $\alpha_0 > 0$ such that

$$\alpha_0 |||u_0|||^2 \leq \mathcal{A}(u_0, u_0), \qquad \forall u_0 \in \mathcal{V}_0,$$

and that the bilinear form $\mathcal{B}(\cdot,\cdot)$ satisfies the following inf-sup condition: there exists $\beta_0 > 0$ such that

$$\beta_0 |||\bar{u}||| \leq \sup_{v_0 \in \mathcal{V}_0} \frac{\mathcal{B}(v_0, \bar{u})}{|||v_0|||} \qquad \forall \bar{u} \in \overline{\mathcal{V}}. \tag{5.3}$$

Then the mixed problem (5.2) has a unique solution $(u_0, \bar{u})$ and the following estimates hold

$$\begin{aligned} |||u_0||| &\leq \frac{1}{\alpha_0} \|\mathcal{L}\|_{\mathcal{V}_0'} + \frac{1}{\beta_0} \left(1 + \frac{\alpha_1}{\alpha_0}\right) \|\mathcal{L}\|_{\overline{\mathcal{V}}'} \\ |||\bar{u}||| &\leq \frac{1}{\beta_0} \left(1 + \frac{\alpha_1}{\alpha_0}\right) \|\mathcal{L}\|_{\mathcal{V}_0'} + \frac{\alpha_1}{\beta_0^2} \left(1 + \frac{\alpha_1}{\alpha_0}\right) \|\mathcal{L}\|_{\overline{\mathcal{V}}'} \end{aligned} \tag{5.4}$$

where

$$\|\mathcal{L}\|_{\mathcal{V}_0'} = \sup_{v_0 \in \mathcal{V}_0} \frac{\mathcal{L}(v_0)}{|||v_0|||} \qquad \textit{and} \qquad \|\mathcal{L}\|_{\overline{\mathcal{V}}'} = \sup_{\bar{v} \in \overline{\mathcal{V}}} \frac{\mathcal{L}(\bar{v})}{|||\bar{v}|||}.$$

Corollary 2. *Under the assumptions of Theorem 6, the bilinear form $\mathcal{A}(\cdot,\cdot)$ is*

bounded on the space $\mathcal{V}$*:*

$$\mathcal{A}(u,v) \leq 2\max\{\alpha_1,\beta_1\} \left|\!\left|\!\left| u \right|\!\right|\!\right| \left|\!\left|\!\left| v \right|\!\right|\!\right|, \qquad \forall u,v \in \mathcal{V}$$

and satisfies the inf-sup condition:

$$c \left|\!\left|\!\left| u \right|\!\right|\!\right| \leq \sup_{v\in\mathcal{V}} \frac{\mathcal{A}(u,v)}{\left|\!\left|\!\left| v \right|\!\right|\!\right|} \qquad \forall u \in \mathcal{V}$$

with $c = \left(\gamma_1^2 + 2\gamma_2^2 + \gamma_3^2\right)^{-1/2}$ *where* γ_i *are the constants from the stability estimate* (5.4)*:*

$$\gamma_1 = \frac{1}{\alpha_0} \qquad \gamma_2 = \frac{1}{\beta_0}\left(1+\frac{\alpha_1}{\alpha_0}\right) \qquad \gamma_3 = \frac{\alpha_1}{\beta_0^2}\left(1+\frac{\alpha_1}{\alpha_0}\right).$$

Proof. The boundedness follows easily from the assumptions: let $u = u_0+\bar{u}$, $v = v_0+\bar{v}$ with $u_0, v_0 \in \mathcal{V}_0$ and $\bar{u}, \bar{v} \in \overline{\mathcal{V}}$ then

$$\begin{aligned}
\mathcal{A}(u,v) &= \mathcal{A}(u_0,v_0) + \mathcal{B}(v_0,\bar{u}) - \mathcal{B}(u_0,\bar{v}) \\
&\leq \alpha_1 \left|\!\left|\!\left| u_0 \right|\!\right|\!\right| \left|\!\left|\!\left| v_0 \right|\!\right|\!\right| + \beta_1 \left|\!\left|\!\left| \bar{u} \right|\!\right|\!\right| \left|\!\left|\!\left| v_0 \right|\!\right|\!\right| + \beta_1 \left|\!\left|\!\left| u_0 \right|\!\right|\!\right| \left|\!\left|\!\left| \bar{v} \right|\!\right|\!\right| \\
&\leq \max\{\alpha_1,\beta_1\} \left(2\left|\!\left|\!\left| u_0 \right|\!\right|\!\right|^2 + \left|\!\left|\!\left| \bar{u} \right|\!\right|\!\right|^2\right)^{\frac{1}{2}} \left(2\left|\!\left|\!\left| v_0 \right|\!\right|\!\right|^2 + \left|\!\left|\!\left| \bar{v} \right|\!\right|\!\right|^2\right)^{\frac{1}{2}} \\
&\leq 2\max\{\alpha_1,\beta_1\} \left|\!\left|\!\left| u \right|\!\right|\!\right| \left|\!\left|\!\left| v \right|\!\right|\!\right|.
\end{aligned}$$

The inf-sup condition follows from the stability estimates (5.4). Let $u \in \mathcal{V}$, $u = u_0+\bar{u}$ with $u_0 \in \mathcal{V}_0$, $\bar{u} \in \overline{\mathcal{V}}$ and define

$$\mathcal{L}_u(v) = \mathcal{A}(u,v), \qquad \forall v \in \mathcal{V}.$$

It is clear that the pair $(u_0, \bar{u})$ is the unique solution to (5.2) with $\mathcal{L} = \mathcal{L}_u$ and therefore the estimates (5.4) hold:

$$\begin{aligned}
\left|\!\left|\!\left| u_0 \right|\!\right|\!\right|^2 &\leq \left(\gamma_1^2+\gamma_2^2\right)\left(\|\mathcal{L}_u\|_{\mathcal{V}_0'}^2 + \|\mathcal{L}_u\|_{\overline{\mathcal{V}}'}^2\right) \\
\left|\!\left|\!\left| \bar{u} \right|\!\right|\!\right|^2 &\leq \left(\gamma_2^2+\gamma_3^2\right)\left(\|\mathcal{L}_u\|_{\mathcal{V}_0'}^2 + \|\mathcal{L}_u\|_{\overline{\mathcal{V}}'}^2\right).
\end{aligned}$$

Adding these two inequalities gives

$$|||u|||^2 = |||u_0|||^2 + |||\bar{u}|||^2 \leq \left(\gamma_1^2 + 2\gamma_2^2 + \gamma_3^2\right)\left(\|\mathcal{L}_u\|_{\mathcal{V}_0'}^2 + \|\mathcal{L}_u\|_{\overline{\mathcal{V}}'}^2\right).$$

Notice that

$$\|\mathcal{L}_u\|_{\mathcal{V}'} = \sup_{v\in\mathcal{V}} \frac{\mathcal{L}_u(v)}{|||v|||} = \sup_{v\in\mathcal{V}} \frac{\mathcal{A}(u,v)}{|||v|||}$$

and therefore to complete the proof we need to show that

$$\|\mathcal{L}_u\|_{\mathcal{V}'}^2 = \|\mathcal{L}_u\|_{\mathcal{V}_0'}^2 + \|\mathcal{L}_u\|_{\overline{\mathcal{V}}'}^2. \tag{5.5}$$

This equality follows from the fact that $\mathcal{V}_0$ and $\overline{\mathcal{V}}$ are orthogonal with respect to the inner product $\langle\!\langle \cdot,\cdot \rangle\!\rangle$. Indeed, if we define $\ell \in \mathcal{V}$, $\ell_0 \in \mathcal{V}_0$, and $\bar{\ell} \in \overline{\mathcal{V}}$ by

$$\begin{aligned}
\langle\!\langle \ell, v \rangle\!\rangle &= \mathcal{L}_u(v), && \forall v \in \mathcal{V},\\
\langle\!\langle \ell_0, v_0 \rangle\!\rangle &= \mathcal{L}_u(v_0), && \forall v_0 \in \mathcal{V}_0,\\
\left\langle\!\left\langle \bar{\ell}, \bar{v} \right\rangle\!\right\rangle &= \mathcal{L}_u(\bar{v}), && \forall \bar{v} \in \overline{\mathcal{V}},
\end{aligned}$$

then for any $v = v_0 + \bar{v}$, $v_0 \in \mathcal{V}_0$, $\bar{v} \in \overline{\mathcal{V}}$ using the orthogonality we have

$$\begin{aligned}
\langle\!\langle \ell, v \rangle\!\rangle = \mathcal{L}_u(v) = \mathcal{L}_u(v_0) + \mathcal{L}_u(\bar{v}) = \langle\!\langle \ell_0, v_0 \rangle\!\rangle + \left\langle\!\left\langle \bar{\ell}, \bar{v} \right\rangle\!\right\rangle\\
= \langle\!\langle \ell_0, v \rangle\!\rangle + \left\langle\!\left\langle \bar{\ell}, v \right\rangle\!\right\rangle = \left\langle\!\left\langle \ell_0 + \bar{\ell}, v \right\rangle\!\right\rangle
\end{aligned}$$

that is $\ell = \ell_0 + \bar{\ell}$. Therefore $|||\ell|||^2 = |||\ell_0|||^2 + \left|||\bar{\ell}|||\right|^2$ and noticing that

$$\|\mathcal{L}_u\|_{\mathcal{V}'} = |||\ell||| \qquad \|\mathcal{L}_u\|_{\mathcal{V}_0'} = |||\ell_0||| \qquad \|\mathcal{L}_u\|_{\overline{\mathcal{V}}'} = \left|||\bar{\ell}|||\right|$$

we obtain (5.5). □

Our next goal is to verify that the assumptions of the above theorem and Corollary hold with constants independent of the element sizes. We begin with the following

Lemma 7. *The non-symmetric bilinear form $\mathcal{A}(\cdot,\cdot)$ is bounded and coercive on the*

discrete space $\mathcal{V}_0$ *with respect to the* $|||\cdot|||$ *norm and with constants in the bounds independent of* $\{h_T\}$.

Proof. The coercivity is easy to see because for any $u_0 \in \mathcal{V}_0$ we have

$$\mathcal{A}(u_0, u_0) = (a\nabla u_0, \nabla u_0) = |||u_0|||^2 .$$

To prove the boundedness it is sufficient to show that

$$\langle \{a\nabla v_0 \cdot \mathbf{n}\}, [\![u_0]\!] \rangle_{\mathcal{E}_i \cup \mathcal{E}_D} \leq C \, |||v_0||| \, |||u_0||| , \qquad \forall u_0, v_0 \in \mathcal{V}_0.$$

To this end, consider a face $F \in \mathcal{E}$ and let $T_i, T_j \in \mathcal{T}_F$ be two elements for which F is a common face (T_i and T_j may be the same element too). Also, let v_i be the restriction of v_0 to T_i and u_j — the restriction of u_0 to T_j. Then we have

$$\begin{aligned} \int_F (a\nabla v_i \cdot \mathbf{n}) u_j &\leq \|a\nabla v_i \cdot \mathbf{n}\|_{0,F} \|u_j\|_{0,F} \leq C \|\nabla v_i\|_{0,F} \|u_j\|_{0,F} \\ &\leq C h_{T_j}^{-1} \|\nabla v_i\|_{0,T_i} \|u_j\|_{0,T_j}. \end{aligned}$$

For the last estimate we used Lemma 1 and the fact that $h_{T_i} \simeq h_{\mathcal{E}}|_F \simeq h_{T_j}$. Using the fact that u_j has zero average over T_j we get

$$h_{T_j}^{-1} \|u_j\|_{0,T_j} \leq C \|\nabla u_j\|_{0,T_j}$$

and thus the above estimate becomes

$$\int_F (a\nabla v_i \cdot \mathbf{n}) u_j \leq C \|\nabla v_i\|_{0,T_i} \|\nabla u_j\|_{0,T_j}.$$

Since the term $\langle \{a\nabla v_0 \cdot \mathbf{n}\}, [\![u_0]\!] \rangle_{\mathcal{E}_i \cup \mathcal{E}_D}$ is a sum of terms like the left hand side of the last estimate (with a coefficient $\pm 1/2$ or 1), we obtain

$$\langle \{a\nabla v_0 \cdot \mathbf{n}\}, [\![u_0]\!] \rangle_{\mathcal{E}_i \cup \mathcal{E}_D} \leq C \, (a\nabla v_0, \nabla v_0)^{\frac{1}{2}} \, (a\nabla u_0, \nabla u_0)^{\frac{1}{2}} = C \, |||v_0||| \, |||u_0|||$$

which implies the boundedness of $\mathcal{A}(\cdot, \cdot)$. □

Lemma 8. *The bilinear form $\mathcal{B}(\cdot,\cdot)$ is bounded (with C independent of $\{h_T\}$):*

$$\mathcal{B}(v_0,\bar{u}) \leq \frac{C}{\sqrt{\kappa}} |||v_0||| \, |||\bar{u}||| \,, \qquad \forall v_0 \in \mathcal{V}_0, \forall \bar{u} \in \overline{\mathcal{V}}.$$

Proof. Let $T \in \mathcal{T}$ be an element and $F \in \mathcal{E}$ be one of its faces. Then using the estimates in the proof of Lemma 2 we have

$$\int_F (a\nabla v_0 \cdot \mathbf{n}) \, [\![\bar{u}]\!] \leq \frac{C}{\sqrt{\kappa}} (a\nabla v_0, \nabla v_0)_T^{\frac{1}{2}} \left\langle \kappa (a\mathbf{n}\cdot\mathbf{n}) h_{\mathcal{E}}^{-1} [\![\bar{u}]\!], [\![\bar{u}]\!] \right\rangle_F^{\frac{1}{2}}.$$

If F is an interior face and we denote the union of the two elements that share the face F by $S = \cup \mathcal{T}_F$ then using the above estimate we can write

$$\int_F \{a\nabla v_0 \cdot \mathbf{n}\} \, [\![\bar{u}]\!] \leq \frac{C}{\sqrt{\kappa}} (a\nabla v_0, \nabla v_0)_S^{\frac{1}{2}} \left\langle \kappa a_{\mathcal{E}} h_{\mathcal{E}}^{-1} [\![\bar{u}]\!], [\![\bar{u}]\!] \right\rangle_F^{\frac{1}{2}}.$$

Summation over $F \in \mathcal{E}_i \cup \mathcal{E}_D$ gives

$$\mathcal{B}(v_0,\bar{u}) = \langle \{a\nabla v_0 \cdot \mathbf{n}\}, [\![\bar{u}]\!] \rangle_{\mathcal{E}_i \cup \mathcal{E}_D} \leq \frac{C}{\sqrt{\kappa}} (a\nabla v_0, \nabla v_0)^{\frac{1}{2}} \left\langle \kappa a_{\mathcal{E}} h_{\mathcal{E}}^{-1} [\![\bar{u}]\!], [\![\bar{u}]\!] \right\rangle_{\mathcal{E}_i \cup \mathcal{E}_D}^{\frac{1}{2}}$$

which is the desired boundedness. □

Remark 9. *In the last two Lemmas 7 and 8 we do not need to assume that the mesh is globally quasi-uniform and therefore the estimates are valid on locally refined meshes as long as the elements are shape regular (non-degenerate). Under the assumptions of Lemma 2 on the coefficient a, the estimate of Lemma 8 is independent of any jumps the coefficient has across interior faces. On the other hand, the boundedness estimate in Lemma 7 does not seem to hold independently of such jumps.*

The last of the assumptions of Theorem 6 we need to verify is the inf-sup condition (5.3) for the bilinear form $\mathcal{B}(\cdot,\cdot)$. We will consider two cases: 1) quadratic and higher order elements, i. e. $r \geq 2$, and 2) linear elements, $r = 1$.

5.2. Quadratic and Higher Order Elements

We begin with the following simple observation:

Lemma 9. *The lowest order Raviart-Thomas finite element space, RT_0, is contained in the space of the gradients of quadratic polynomials, ∇P_2,*

$$RT_0 \subset \nabla P_2.$$

Proof. First we consider the 2D case (triangle elements). The functions in RT_0 have the form

$$\begin{pmatrix} ax+b \\ ay+c \end{pmatrix}$$

which is the gradient of

$$(a/2)(x^2+y^2)+bx+cy.$$

In the 3D case (tetrahedral elements) RT_0 functions have the form

$$\begin{pmatrix} ax+b \\ ay+c \\ az+d \end{pmatrix}$$

which is the gradient of

$$(a/2)(x^2+y^2+z^2)+bx+cy+dz.$$

□

Lemma 10. *Assume that the coefficient a is scalar and piecewise constant with respect to the mesh $\mathcal{T}$. Then the following inf-sup condition holds for simplicial meshes (i.e. meshes build from triangles or tetrahedra) and for quadratic and higher order*

elements, $r \geq 2$

$$|||\bar{u}||| \leq C\sqrt{\kappa} \sup_{v_0 \in \mathcal{V}_0} \frac{\mathcal{B}(v_0, \bar{u})}{|||v_0|||}, \qquad \forall \bar{u} \in \overline{\mathcal{V}}$$

with C independent of $\{h_T\}$, the coefficient a, and κ.

Proof. Denote the lowest order Raviart-Thomas space on the mesh $\mathcal{T}$ by

$$RT_0(\mathcal{T}) = \{\mathbf{w} \in H(\mathrm{div}; \Omega) : \mathbf{w}|_T \in RT_0, \ \forall T \in \mathcal{T}\}.$$

The degrees of freedom of $\mathbf{w} \in RT_0(\mathcal{T})$ are the integrals $\int_F \mathbf{w} \cdot \mathbf{n}$ for all faces $F \in \mathcal{E}$. Note that $[\![\mathbf{w} \cdot \mathbf{n}]\!]_F = 0$ for all interior faces $F \in \mathcal{E}_i$ and that $(\mathbf{w} \cdot \mathbf{n})|_F$ is constant. Take a fixed $\bar{u} \in \overline{\mathcal{V}}$ and define the vector function $\mathbf{w} \in RT_0(\mathcal{T})$ by

$$(\mathbf{w} \cdot \mathbf{n})|_F = \begin{cases} (h_{\mathcal{E}}^{-1} [\![\bar{u}]\!])|_F & \forall F \in \mathcal{E}_i \cup \mathcal{E}_D \\ 0 & \forall F \in \mathcal{E}_N. \end{cases}$$

Let T be an element and let $\hat{\mathbf{w}}$ denote the mapping of $\mathbf{w}|_T$ back to the reference element $\hat{T}$ via Piola's transformation. The following estimate holds (see [11], page 98)

$$\|\mathbf{w}\|_{0,T}^2 \leq C h_T^{2-d} \|\hat{\mathbf{w}}\|_{0,\hat{T}}^2.$$

On the reference element the term $\|\hat{\mathbf{w}}\|_{0,\hat{T}}^2$ is equivalent to the sum of the squares of the degrees of freedom

$$\|\hat{\mathbf{w}}\|_{0,\hat{T}}^2 \simeq \sum_{\hat{F}} \left(\int_{\hat{F}} \hat{\mathbf{w}} \cdot \hat{\mathbf{n}} \right)^2$$

where the sum is over the faces $\hat{F}$ of $\hat{T}$. Since the degrees of freedom are preserved under the Piola transformation (and the direction of $\mathbf{n}$ only changes the sign) we get

$$\|\mathbf{w}\|_{0,T}^2 \leq C h_T^{2-d} \sum_{F \in \mathcal{E}_T} \left(\int_F \mathbf{w} \cdot \mathbf{n} \right)^2 \leq C \sum_{F \in \mathcal{E}_T} h_{\mathcal{E}}^{2-d} h_{\mathcal{E}}^{d-1} \int_F (\mathbf{w} \cdot \mathbf{n})^2$$

where $\mathcal{E}_T$ denotes the set of all faces of T and we used the fact that $(\mathbf{w} \cdot \mathbf{n})$ is constant

over each face. Multiplying by the constant coefficient a, summing over all elements T, and using the definition of $(\mathbf{w}\cdot\mathbf{n})$ we obtain the estimate

$$(a\mathbf{w},\mathbf{w}) \le C \sum_{F\in\mathcal{E}_i\cup\mathcal{E}_D} \int_F \{a\}\, h_{\mathcal{E}}^{-1} \left[\!\left[\bar{u}\right]\!\right]^2 = \frac{C}{\kappa} \left\langle \kappa a_{\mathcal{E}} h_{\mathcal{E}}^{-1} \left[\!\left[\bar{u}\right]\!\right], \left[\!\left[\bar{u}\right]\!\right] \right\rangle_{\mathcal{E}_i\cup\mathcal{E}_D} = \frac{C}{\kappa} \left|\!\left|\!\left|\bar{u}\right|\!\right|\!\right|^2. \quad (5.6)$$

Now, using Lemma 9 we can construct a piecewise quadratic function z satisfying

$$\nabla z|_T = \mathbf{w}|_T \quad \text{and} \quad \int_T z = 0, \qquad \forall T\in\mathcal{T},$$

so that $z\in\mathcal{V}_0$. Note that the same piecewise quadratic z is used for all polynomial degrees $r\ge 2$. In addition, since $\mathbf{w}$ has continuous normal components across interior faces, we have that

$$\{a\nabla z\cdot\mathbf{n}\}\,|_F = \{a\}\,(\nabla z\cdot\mathbf{n})|_F = \{a\}\,(\mathbf{w}\cdot\mathbf{n})|_F = (a_{\mathcal{E}} h_{\mathcal{E}}^{-1} \left[\!\left[\bar{u}\right]\!\right])|_F, \qquad \forall F\in\mathcal{E}_i\cup\mathcal{E}_D.$$

Also, using (5.6) we have that

$$\left|\!\left|\!\left|z\right|\!\right|\!\right|^2 = (a\nabla z,\nabla z) = (a\mathbf{w},\mathbf{w}) \le \frac{C}{\kappa} \left|\!\left|\!\left|\bar{u}\right|\!\right|\!\right|^2.$$

Combining the above equality for $\{a\nabla z\cdot\mathbf{n}\}$ and last estimate we obtain

$$\begin{aligned} \left|\!\left|\!\left|\bar{u}\right|\!\right|\!\right| &= \frac{\kappa \left\langle a_{\mathcal{E}} h_{\mathcal{E}}^{-1} \left[\!\left[\bar{u}\right]\!\right], \left[\!\left[\bar{u}\right]\!\right] \right\rangle_{\mathcal{E}_i\cup\mathcal{E}_D}}{\left|\!\left|\!\left|\bar{u}\right|\!\right|\!\right|} = \frac{\kappa \left\langle \{a\nabla z\cdot\mathbf{n}\}, \left[\!\left[\bar{u}\right]\!\right] \right\rangle_{\mathcal{E}_i\cup\mathcal{E}_D}}{\left|\!\left|\!\left|\bar{u}\right|\!\right|\!\right|} = \frac{\kappa\mathcal{B}(z,\bar{u})}{\left|\!\left|\!\left|\bar{u}\right|\!\right|\!\right|} \\ &\le C\sqrt{\kappa}\frac{\mathcal{B}(z,\bar{u})}{\left|\!\left|\!\left|z\right|\!\right|\!\right|} \le C\sqrt{\kappa} \sup_{v_0\in\mathcal{V}_0} \frac{\mathcal{B}(v_0,\bar{u})}{\left|\!\left|\!\left|v_0\right|\!\right|\!\right|}. \end{aligned}$$

□

5.3. Linear Elements

In this section we will study the properties of the bilinear form $\mathcal{B}(\cdot,\cdot)$ in the case of linear elements, $r=1$. Introduce the linear operator $B:\mathcal{V}_0\to\overline{\mathcal{V}}'$ and its transpose

$B^t : \overline{\mathcal{V}} \to \mathcal{V}_0'$ by

$$(Bv_0)(\bar{v}) = (B^t\bar{v})(v_0) = \mathcal{B}(v_0, \bar{v}), \quad \forall v_0 \in \mathcal{V}_0,\ \forall \bar{v} \in \overline{\mathcal{V}}.$$

Note that the inf-sup condition (5.3) is equivalent to the estimate

$$c\,|||\bar{u}||| \le \|B^t\bar{u}\|_{\mathcal{V}_0'}, \qquad \forall \bar{u} \in \overline{\mathcal{V}}$$

and therefore a necessary condition for the inf-sup condition is that the kernel of the operator B^t is trivial: $\operatorname{Ker} B^t = \{0\}$. We will show that for certain types of meshes $\operatorname{Ker} B^t$ is not trivial when linear elements are used and therefore the inf-sup condition (5.3) does not hold in that case.

In this section we will assume that the coefficient $a = 1$. We begin with the following auxiliary

Lemma 11. *Let T be a non-degenerate d-simplex with faces $\{F_i\}_{i=1}^{d+1}$, and let $\{\mathbf{n}_i\}_{i=1}^{d+1}$ be the outward normals to the faces. Then*

$$\sum_{i=1}^{d+1} \alpha_i |F_i| \mathbf{n}_i = \mathbf{0}, \qquad (\alpha_i \in \mathbb{R})$$

if and only if $\alpha_i = \alpha_j,\ \forall i, j$.

Proof. Let $\mathbf{w} \in \mathbb{R}^d$ be arbitrary, then using the divergence theorem we get

$$\mathbf{w} \cdot \left(\sum_{i=1}^{d+1} |F_i| \mathbf{n}_i \right) = \mathbf{w} \cdot \left(\int_{\partial T} \mathbf{n} \right) = \int_{\partial T} \mathbf{w} \cdot \mathbf{n} = \int_T \operatorname{div} \mathbf{w} = 0$$

which implies that

$$\sum_{i=1}^{d+1} |F_i| \mathbf{n}_i = \mathbf{0}.$$

Since any d of the normal vectors $\{\mathbf{n}_i\}$ are linearly independent the above equality completes the proof. □

Definition 1. *We will call the mesh $\mathcal{T}$ a checkerboard mesh if it satisfies the following*

property: there exists a coloring of the elements with two colors such that the neighbors of every element have a color different from the color of the element itself. (Two elements are neighbors if they have a common $(d-1)$-dimensional face).

Definition 2. *We will call the mesh $\mathcal{T}$ connected if there exists a path between every two elements of the mesh. A path between T and T' is a sequence of elements $T_1, T_2, \ldots, T_n$ such that $T_1 = T$, $T_n = T'$ and $\{T_i, T_{i+1}\}$ are neighbors for all $i = 1, \ldots, n-1$.*

The following lemma gives a characterization of $\operatorname{Ker} B^t$.

Lemma 12. *Let $\mathcal{T}$ be a connected simplicial mesh and $r = 1$ (linear elements). Then $\operatorname{Ker} B^t$ is non-trivial in the following two cases:*

- *The mesh $\mathcal{T}$ is a checkerboard mesh and $\Gamma_D = \partial\Omega$. In this case $\operatorname{Ker} B^t$ is one-dimensional and is spanned by the checkerboard function defined as $+1$ at the elements with one of the colors (from the definition of checkerboard mesh) and -1 at the elements with the other color.*
- *We have $\Gamma_N = \partial\Omega$. In this case $\operatorname{Ker} B^t$ is the one-dimensional space of all constant functions.*

In all other cases $\operatorname{Ker} B^t = \{0\}$.

Proof. Let $v_0 \in \mathcal{V}_0$, $\bar{u} \in \overline{\mathcal{V}}$. Consider an interior face $F \in \mathcal{E}_i$ and let T_1 and T_2 be the two elements that share the face F ($T_1 \neq T_2$). Denote $v_i = v_0|_{T_i}$, $u_i = \bar{u}|_{T_i}$, $i = 1, 2$ and let $\mathbf{n}_i$ be the vector normal to F pointing outside of T_i, $i = 1, 2$. Assume that $\mathbf{n}|_F = \mathbf{n}_1$ then we have

$$\begin{aligned}(\{\nabla v_0 \cdot \mathbf{n}\}\,[\![\bar{u}]\!])|_F &= \frac{1}{2}(\nabla v_1 \cdot \mathbf{n} + \nabla v_2 \cdot \mathbf{n})(u_1 - u_2) \\ &= (\nabla v_1 \cdot \mathbf{n}_1)\frac{1}{2}(u_1 - u_2) + (\nabla v_2 \cdot \mathbf{n}_2)\frac{1}{2}(u_2 - u_1)\end{aligned}$$

Thus, if we define the double valued function $[\bar{u}]$ on F by

$$([\bar{u}]|_{\partial T_1})|_F = \frac{1}{2}(u_1 - u_2) = \frac{1}{2}(\bar{u}|_{T_1} - \bar{u}|_{T_2})$$
$$([\bar{u}]|_{\partial T_2})|_F = \frac{1}{2}(u_2 - u_1) = \frac{1}{2}(\bar{u}|_{T_2} - \bar{u}|_{T_1})$$

we can rewrite the bilinear form $\mathcal{B}(\cdot,\cdot)$ as

$$\begin{aligned}\mathcal{B}(v_0,\bar{u}) &= \sum_{F\in\mathcal{E}_i}\int_F \{\nabla v_0\cdot\mathbf{n}\}\,[\![\bar{u}]\!] + \sum_{F\in\mathcal{E}_D}\int_F (\nabla v_0\cdot\mathbf{n})\bar{u}\\ &= \sum_{T\in\mathcal{T}}\int_{\partial T}(\nabla v_0\cdot\mathbf{n}_T)[\bar{u}]\end{aligned}$$

where $\mathbf{n}_T = \pm\mathbf{n}$ is the normal vector pointing outside of T and we extended the definition of $[\bar{u}]$ to $\partial\Omega$ by

$$[\bar{u}]|_F = \begin{cases}\bar{u}, & F\in\mathcal{E}_D\\ 0, & F\in\mathcal{E}_N.\end{cases}$$

Observing that ∇v_0 is constant over the elements we can write

$$\mathcal{B}(v_0,\bar{u}) = \sum_{T\in\mathcal{T}}\nabla v_0\cdot\int_{\partial T}\mathbf{n}_T[\bar{u}]. \tag{5.7}$$

We want to show that

$$\operatorname{Ker} B^t = \left\{\bar{u}\in\overline{\mathcal{V}} : \int_{\partial T}\mathbf{n}_T[\bar{u}] = \mathbf{0},\ \forall T\in\mathcal{T}\right\}. \tag{5.8}$$

Assume that $\bar{u}\in\operatorname{Ker} B^t$ then $\mathcal{B}(v_0,\bar{u}) = 0$, $\forall v_0\in\mathcal{V}_0$. In particular, we can take v_0 such that

$$\nabla v_0|_T = \int_{\partial T}\mathbf{n}_T[\bar{u}]$$

and therefore

$$0 = \mathcal{B}(v_0,\bar{u}) = \sum_{T\in\mathcal{T}}\left|\int_{\partial T}\mathbf{n}_T[\bar{u}]\right|^2$$

which proves the "$\subseteq$" part of (5.8). The other direction "$\supseteq$" of (5.8) is easy to see

given (5.7). With the help of Lemma 11 the condition $\int_{\partial T} \mathbf{n}_T[\bar{u}] = \mathbf{0}$ can be replaced by $[\bar{v}]|_{\partial T} = \text{const}$ and thus we have the characterization

$$\operatorname{Ker} B^t = \left\{ \bar{u} \in \overline{\mathcal{V}} : [\bar{u}]|_{\partial T} = \text{const},\ \forall T \in \mathcal{T} \right\}.$$

Using the definition of $[\bar{u}]$ the condition $[\bar{u}]|_{\partial T} = \text{const}$ can be interpreted as follows:

- If $T \in \mathcal{T}$ is arbitrary, then $\bar{u}$ has the same value at all neighbor elements of T.
- If T is such that $\mathcal{E}_T \cap \mathcal{E}_D \neq \emptyset$ (recall that $\mathcal{E}_T$ denotes the set of all faces of T), then the value of $\bar{u}$ at the neighbor elements of T is minus the value of $\bar{u}$ at T.
- If T is such that $\mathcal{E}_T \cap \mathcal{E}_N \neq \emptyset$, then the value of $\bar{u}$ at the neighbor elements of T is the same as the value of $\bar{u}$ at T.

Let $\bar{u} \in \operatorname{Ker} B^t$ and let $T_0 \in \mathcal{T}$ be a fixed boundary element (i. e. $\mathcal{E}_{T_0} \cap \mathcal{E}_b \neq \emptyset$). Denote the value of $\bar{u}$ at T_0 by $\alpha = \bar{u}|_{T_0}$ and let $T \in \mathcal{T}$ be an arbitrary element. Since the mesh is connected there exists a path from T_0 to T: $T_0, T_1, \ldots, T_n = T$. Using the above three properties of the functions in $\operatorname{Ker} B^t$ we can conclude that

$$\begin{aligned} &\bar{u}|_{T_i} = (-1)^i \alpha, \quad i = 1, \ldots, n \qquad \text{if} \quad \mathcal{E}_{T_0} \cap \mathcal{E}_D \neq \emptyset \\ &\bar{u}|_{T_i} = \alpha, \qquad\quad i = 1, \ldots, n \qquad \text{if} \quad \mathcal{E}_{T_0} \cap \mathcal{E}_N \neq \emptyset. \end{aligned} \tag{5.9}$$

Thus, knowing the value of $\bar{u}$ on the fixed boundary element T_0 allows us to recover the function everywhere. This shows that $\operatorname{Ker} B^t$ has dimension at most 1.

If $\mathcal{T}$ is checkerboard mesh and $\Gamma_D = \partial\Omega$ then the checkerboard function $\bar{u}$ satisfies $[\bar{u}]|_{\partial T} = 1$ when $\bar{u}|_T = 1$ and $[\bar{u}]|_{\partial T} = -1$ when $\bar{u}|_T = -1$ and therefore $\bar{u} \in \operatorname{Ker} B^t$. Since the dimension of $\operatorname{Ker} B^t$ is at most 1, we have $\operatorname{Ker} B^t = \operatorname{span}\{\bar{u}\}$.

If $\Gamma_N = \partial\Omega$ then the constant function $\bar{u} \equiv 1$ satisfies $[\bar{u}]|_{\partial T} = 0$ for all $T \in \mathcal{T}$. As in the previous case this means that $\operatorname{Ker} B^t = \operatorname{span}\{\bar{u}\}$.

To prove that $\operatorname{Ker} B^t = \{0\}$ in all other cases we will show that if $\operatorname{Ker} B^t$ is

non-trivial then we have one of the two cases above. Let $\bar{u} \in \operatorname{Ker} B^t$, $\bar{u} \neq 0$. We want to show that either $\mathcal{E}_D = \emptyset$ or $\mathcal{E}_N = \emptyset$. If $\mathcal{E}_N \neq \emptyset$ then the second equality in (5.9) implies that $\bar{u} \equiv \text{const}$ and therefore $[\bar{u}]|_{\partial T} = 0$ for all $T \in \mathcal{T}$. If we also assume that $\mathcal{E}_D \neq \emptyset$ then for any $F \in \mathcal{E}_D$ we have $0 = [\bar{u}]|_F = \bar{u}|_F$ which is a contradiction. Thus, either $\Gamma_N = \partial\Omega$ or $\Gamma_D = \partial\Omega$. In the latter case we need to show that $\mathcal{T}$ is a checkerboard mesh. This follows from the first equality in (5.9) since $\alpha \neq 0$. □

Remark 10. *The kernels of the operators $A : \mathcal{V} \to \mathcal{V}'$ and $A^t : \mathcal{V} \to \mathcal{V}'$ defined by*

$$(Au)(v) = (A^t v)(u) = \mathcal{A}(u, v), \qquad \forall u, v \in \mathcal{V}$$

coincide with the kernel of B^t: $\operatorname{Ker} A = \operatorname{Ker} A^t = \operatorname{Ker} B^t$.

Proof. Let $u \in \operatorname{Ker} A$ then

$$0 = (Au)(u) = \mathcal{A}(u, u) = (\nabla u, \nabla u)$$

that is $u \in \overline{\mathcal{V}}$ and therefore

$$0 = (Au)(v_0) = \mathcal{A}(u, v_0) = \mathcal{B}(v_0, u), \qquad \forall v_0 \in \mathcal{V}_0$$

that is $u \in \operatorname{Ker} B^t$. Similarly, if $u \in \operatorname{Ker} A^t$ then $u \in \overline{\mathcal{V}}$ and

$$0 = (A^t u)(v_0) = \mathcal{A}(v_0, u) = -\mathcal{B}(v_0, u), \qquad \forall v_0 \in \mathcal{V}_0$$

that is $u \in \operatorname{Ker} B^t$. Conversely, if $\bar{u} \in \operatorname{Ker} B^t$ then for any $v \in \mathcal{V}$ decomposed as $v = v_0 + \bar{v}$, $v_0 \in \mathcal{V}_0$, $\bar{v} \in \overline{\mathcal{V}}$ we have

$$(A\bar{u})(v) = \mathcal{A}(\bar{u}, v) = \mathcal{B}(v_0, \bar{u}) = 0$$
$$(A^t\bar{u})(v) = \mathcal{A}(v, \bar{u}) = -\mathcal{B}(v_0, \bar{u}) = 0,$$

that is $\bar{u} \in \operatorname{Ker} A \cap \operatorname{Ker} A^t$. This completes the proof. □

In the remainder of the section we will derive an equivalent form of the inf-sup condition (5.3). Let $CR(\mathcal{T})$ be the linear non-conforming Crouzeix-Raviart finite element space on the mesh $\mathcal{T}$:

$$CR(\mathcal{T}) = \left\{ v \in L^2(\Omega) : v|_T \in P_1(T),\ \forall T \in \mathcal{T},\ \text{and}\ \int_F [\![v]\!] = 0,\ \forall F \in \mathcal{E}_i \right\}$$

and define the operator $P : \overline{\mathcal{V}} \to CR(\mathcal{T})$ by: for all $F \in \mathcal{E}$

$$(P\bar{u})(M_F) = \begin{cases} \{\bar{u}\}\,(M_F), & F \in \mathcal{E}_i \cup \mathcal{E}_N \\ 0, & F \in \mathcal{E}_D, \end{cases}$$

where M_F denotes the midpoint of the face F. Using the definitions of P and $[\cdot]$ we can show that the following equality holds for any $\bar{u} \in \overline{\mathcal{V}}$

$$([\bar{u}]|_{\partial T})(M_F) = (\bar{u}|_T)(M_F) - (P\bar{u})(M_F), \qquad \forall T \in \mathcal{T} \text{ and } F \in \mathcal{E}_T. \tag{5.10}$$

Indeed, if $F \in \mathcal{E}_i$ is an interior face shared by the elements T_1 and T_2 and we denote $u_i = (\bar{u}|_{T_i})(M_F)$, $i = 1, 2$ then we have

$$(\bar{u}|_{T_1})(M_F) - (P\bar{u})(M_F) = u_1 - \frac{1}{2}(u_1 + u_2) = \frac{1}{2}(u_1 - u_2) = ([\bar{u}]|_{\partial T_1})(M_F).$$

If $F \in \mathcal{E}_D$ and T is the element for which F is a face, then

$$(\bar{u}|_T)(M_F) - (P\bar{u})(M_F) = (\bar{u}|_T)(M_F) - 0 = ([\bar{u}]|_{\partial T})(M_F).$$

Finally, if T is a boundary element with a face $F \in \mathcal{E}_N$, then

$$(\bar{u}|_T)(M_F) - (P\bar{u})(M_F) = (\bar{u}|_T)(M_F) - (\bar{u}|_T)(M_F) = 0 = ([\bar{u}]|_{\partial T})(M_F).$$

Note that if v is linear on F we have

$$\int_F v = |F|\, v(M_F)$$

and therefore (5.10) implies that for any $\bar{u} \in \overline{\mathcal{V}}$ we have

$$\int_F [\bar{u}]|_{\partial T} = \int_F (\bar{u} - P\bar{u})|_T, \qquad \forall T \in \mathcal{T} \text{ and } F \in \mathcal{E}_T. \tag{5.11}$$

Lemma 13. *The following equalities hold*

$$\sup_{v_0 \in \mathcal{V}_0} \frac{\mathcal{B}(v_0, \bar{u})}{|||v_0|||} = \left(\sum_{T \in \mathcal{T}} |T|^{-1} \left| \int_{\partial T} \mathbf{n}_T[\bar{u}] \right|^2 \right)^{1/2} = \|\nabla(\bar{u} - P\bar{u})\|, \qquad \forall \bar{u} \in \overline{\mathcal{V}}.$$

(Note that we use ∇ to denote the element-by-element gradient.)

Proof. Let us define $z \in \mathcal{V}_0$ by

$$\nabla z|_T = |T|^{-1} \int_{\partial T} \mathbf{n}_T[\bar{u}] \qquad \text{and} \qquad \int_T z = 0, \qquad \forall T \in \mathcal{T}.$$

Then using formula (5.7) we can write for any $v_0 \in \mathcal{V}_0$:

$$\mathcal{B}(v_0, \bar{u}) = \sum_{T \in \mathcal{T}} \nabla v_0 \cdot \int_{\partial T} \mathbf{n}_T[\bar{u}] = \sum_{T \in \mathcal{T}} |T|(\nabla v_0 \cdot \nabla z) = (\nabla v_0, \nabla z)$$

and therefore we have

$$\sup_{v_0 \in \mathcal{V}_0} \frac{\mathcal{B}(v_0, \bar{u})}{|||v_0|||} = \sup_{v_0 \in \mathcal{V}_0} \frac{(\nabla v_0, \nabla z)}{(\nabla v_0, \nabla v_0)^{1/2}} = |||z||| = \left(\sum_{T \in \mathcal{T}} |T|^{-1} \left| \int_{\partial T} \mathbf{n}_T[\bar{u}] \right|^2 \right)^{1/2}$$

which is the first equality we had to prove. To establish the second equality we will show that $\nabla(\bar{u} - P\bar{u}) = \nabla z$. Let $T \in \mathcal{T}$ and $\mathbf{w} \in \mathbb{R}^d$ be arbitrary, then using (5.11), the fact that $(\mathbf{w} \cdot \mathbf{n}_T)$ is constant over each face of T, and the divergence theorem we obtain

$$\begin{aligned} \mathbf{w} \cdot \int_{\partial T} \mathbf{n}_T[\bar{u}] &= \int_{\partial T} (\mathbf{w} \cdot \mathbf{n}_T)[\bar{u}] = \int_{\partial T} (\mathbf{w} \cdot \mathbf{n}_T)(\bar{u} - P\bar{u}) \\ &= \int_T (\operatorname{div} \mathbf{w})(\bar{u} - P\bar{u}) + \int_T \mathbf{w} \cdot \nabla(\bar{u} - P\bar{u}) = \mathbf{w} \cdot \int_T \nabla(\bar{u} - P\bar{u}). \end{aligned}$$

This implies that

$$\int_{\partial T} \mathbf{n}_T[\bar{u}] = \int_T \nabla(\bar{u} - P\bar{u}) = |T| \nabla(\bar{u} - P\bar{u})|_T.$$

Comparing with the definition of z we see that $\nabla z = \nabla(\bar{u} - P\bar{u})$ and therefore

$$\sup_{v_0 \in \mathcal{V}_0} \frac{\mathcal{B}(v_0, \bar{u})}{|||v_0|||} = |||z||| = \|\nabla z\| = \|\nabla(\bar{u} - P\bar{u})\| = \|\nabla P\bar{u}\|.$$

The proof is complete. □

Lemma 14. *The following two-sided estimate holds with constants independent of the element diameters* $\{h_T\}$*:*

$$|||\bar{u}|||^2 \simeq \kappa \sum_{T \in \mathcal{T}} h_T^{-2} \|\bar{u} - P\bar{u}\|_{0,T}^2, \qquad \forall \bar{u} \in \overline{\mathcal{V}}.$$

Proof. Let $\bar{u} \in \overline{\mathcal{V}}$ be arbitrary. Note that the values of $[\bar{u}]|_F$ on the two sides of an interior face F differ only by their sign and therefore $([\bar{u}]|_F)^2$ is single-valued and by the definition of $[\cdot]$ we have $([\![\bar{u}]\!]|_F)^2 = 4([\bar{u}]|_F)^2$ and therefore (for $\kappa = 1$)

$$\begin{aligned}
|||\bar{u}|||^2 &= \sum_{F \in \mathcal{E}_i} \int_F h_{\mathcal{E}}^{-1} [\![\bar{u}]\!]^2 + \sum_{F \in \mathcal{E}_D} \int_F h_{\mathcal{E}}^{-1} \bar{u}^2 = \sum_{F \in \mathcal{E}_i} \int_F h_{\mathcal{E}}^{-1} 4[\bar{u}]^2 + \sum_{F \in \mathcal{E}_D} \int_F h_{\mathcal{E}}^{-1} [\bar{u}]^2 \\
&\simeq \sum_{F \in \mathcal{E}_i} \int_F h_{\mathcal{E}}^{-1} 2[\bar{u}]^2 + \sum_{F \in \mathcal{E}_D} \int_F h_{\mathcal{E}}^{-1} [\bar{u}]^2 = \sum_{T \in \mathcal{T}} \int_{\partial T} h_{\mathcal{E}}^{-1} [\bar{u}]^2 \\
&= \sum_{T \in \mathcal{T}} \sum_{F \in \mathcal{E}_T} (h_{\mathcal{E}}|_F)^{d-2} ([\bar{u}]|_F)^2 \simeq \sum_{T \in \mathcal{T}} h_T^{-2} \|\bar{u} - P\bar{u}\|_{0,T}^2.
\end{aligned}$$

For the last equivalence we used (5.10), the fact that $h_{\mathcal{E}}|_F \simeq h_T$ when $F \in \mathcal{E}_T$, and the following estimate valid for linear $v|_T$

$$\|v\|_{0,T}^2 \simeq h_T^d \sum_{F \in \mathcal{E}_T} (v(M_F))^2$$

which is easily derived by mapping to a reference simplex. □

As a result of the last two lemmas we can write the inf-sup condition (5.3) in the following equivalent (for linear elements) form:

$$c \sum_{T \in \mathcal{T}} h_T^{-2} \|\bar{u} - P\bar{u}\|_{0,T}^2 \leq \|\nabla(\bar{u} - P\bar{u})\|^2 = \|\nabla P\bar{u}\|^2, \qquad \forall \bar{u} \in \overline{\mathcal{V}}. \tag{5.12}$$

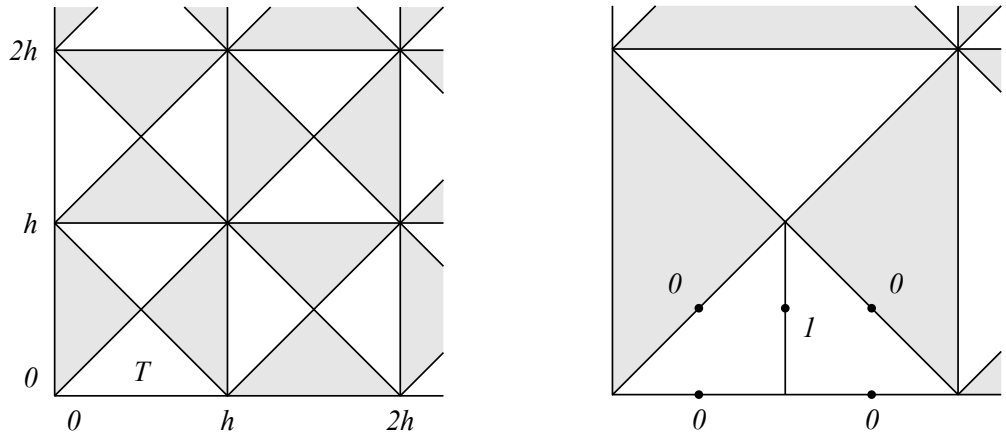

Fig. 5.1. Checkerboard mesh for the unit square (left) and the new mesh obtained after refinement of T (right).

Unfortunately, this estimate does not always hold with constant c independent of h. To show that we consider the following example: let $\Omega = (0,1)^2$ be the unit square, $\Gamma_D = \partial\Omega$, and let $\mathcal{T}_c$ be the checkerboard mesh shown on the left side of Figure 5.1. We define a new mesh $\mathcal{T}$ obtained from $\mathcal{T}_c$ by bisecting the element T as shown on the right side of the figure. Note that the mesh $\mathcal{T}$ is not checkerboard. Let $\bar{z}$ be the checkerboard function on the mesh $\mathcal{T}_c$ that is defined as $+1$ at the white elements and -1 at the gray elements. We define $\mathcal{V}$, $\overline{\mathcal{V}}$, $\mathcal{V}_0$, P, etc. on the mesh $\mathcal{T}$ as before. We have that $P\bar{z}$ vanishes at the midpoints of all boundary edges because $\Gamma_D = \partial\Omega$. Using the definition of $\bar{z}$ it is easy to see that $P\bar{z}$ also vanishes at the midpoints of all interior edges except the one that we used to bisect T as shown on the right side of Figure 5.1. Thus, $P\bar{z}$ is identically zero on all elements in $\mathcal{T}$ except the two elements obtained from bisecting T which we will denote by T_1 and T_2. Moreover, $P\bar{z}$ is equal to the nodal basis function in $CR(\mathcal{T})$ corresponding to the edge shared by T_1 and T_2 and one can easily compute that $\|\nabla P\bar{z}\|^2 = 4$. We have

$$\frac{\|\nabla P\bar{z}\|^2}{\sum_{\tau\in\mathcal{T}} h_\tau^{-2}\|\bar{z} - P\bar{z}\|_{0,\tau}^2} \leq \frac{4}{\sum_{\tau\in\mathcal{T}\setminus\{T_1,T_2\}} h^{-2}\|\bar{z}\|_{0,\tau}^2} = \frac{4h^2}{1-h^2/4} \leq \frac{16}{3}h^2, \quad (h \leq 1)$$

and therefore the constant c in (5.12) cannot exceed $16h^2/3$ and consequently $\beta_0 \lesssim h/\kappa^{\frac{1}{2}}$ where β_0 is the constant in (5.3).

This example shows that the inf-sup condition for linear elements is closely related to the properties of the mesh and some assumptions on the triangulation need to be made (in addition to $\operatorname{Ker} B^t = \{0\}$) in order to be able to prove the estimate independently of the element diameters h_T if that is at all possible.

We conclude this section with a numerical test investigating the convergence rates of the method of Baumann and Oden with linear elements in L^2 and H^1-like norms. We solve the problem $-\Delta u = f$ in the unit cube $\Omega = (0,1)^3$ with Dirichlet boundary conditions imposed on the whole boundary $\Gamma_D = \partial\Omega$ and with the following exact solution $u = -(x^2 + y^2 + z^2)/6$. We use a coarse tetrahedral mesh ("Level 0") that we refine uniformly to obtain a sequence of nested meshes. Thus, the elements in the "Level k" mesh have diameters $h_k = 2^{-k}h_0$. All meshes (except the coarsest "Level 0" mesh) are **checkerboard** and therefore the operators A and A^t have the same kernel which consists of all checkerboard functions (see Lemma 12 and Remark 10). With our choice of the exact solution and the triangulations we have that $\mathcal{L}(v) = 0,\ \forall v \in \operatorname{Ker} A^t$, that is the discrete linear systems are compatible. From the multiple discrete solutions we select the one that is L^2-orthogonal to the checkerboard functions. The results from this numerical experiment are presented in Table 5.1. The top part of the table gives the L^2 and "H^1" norms of the error. We have broken the "H^1" norm into its two components:

$$\text{"grad"}:\ (\nabla v, \nabla v)^{1/2} \qquad \text{and} \qquad \text{"jump"}:\ \left\langle h_{\mathcal{E}}^{-1} [\![v]\!], [\![v]\!] \right\rangle_{\mathcal{E}_i \cup \mathcal{E}_D}^{1/2}.$$

The bottom part of the table gives the corresponding convergence ratios. Clearly, we observe optimal convergence in both L^2 and "H^1" norms — $\mathcal{O}(h^2)$ and $\mathcal{O}(h)$, respectively.

Table 5.1. Convergence of the method of Baumann and Oden with linear elements

	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
L^2, $\times 10^{-6}$	4409.	1068.	257.0	62.34	15.28	3.777
grad, $\times 10^{-4}$	495.9	241.8	118.8	58.84	29.27	14.60
jump, $\times 10^{-4}$	299.8	123.3	50.71	21.68	9.736	4.560
L^2	—	4.129	4.154	4.122	4.080	4.046
grad	—	2.051	2.034	2.019	2.010	2.005
jump	—	2.431	2.432	2.339	2.227	2.135

5.4. Preconditioning

Earlier in this chapter we proved that the inf-sup condition and boundedness

$$c\,|||u||| \leq \sup_{v\in\mathcal{V}} \frac{\mathcal{A}(u,v)}{|||v|||} \leq C\,|||u|||\,, \qquad \forall u \in \mathcal{V} \tag{5.13}$$

hold when $\mathcal{V}$ is built from quadratic or higher order elements ($r \geq 2$). In this section we will show how this estimates can be used to define preconditioners for the discrete problem (5.1). Let $\{\phi_i\}_{i=1}^n$ be a basis for $\mathcal{V}$ and let us define the matrices A and B with entries given by

$$A_{ij} = \mathcal{A}(\phi_j, \phi_i) \qquad \text{and} \qquad B_{ij} = \langle\langle \phi_j, \phi_i \rangle\rangle\,, \qquad i,j = 1,\ldots,n.$$

Note that B is symmetric and positive definite and A is non-symmetric and its entries are given by $\mathcal{A}(\phi_j, \phi_i)$ and not by $\mathcal{A}(\phi_i, \phi_j)$. If we define the column vectors x and b with entries

$$x_i : u = \sum_{i=1}^{n} x_i \phi_i \qquad \text{and} \qquad b_i = \mathcal{L}(\phi_i), \quad i = 1,\ldots,n$$

then the discrete problem (5.1) can be written as

$$Ax = b.$$

If we also define the column vector y with entries

$$y_i : v = \sum_{i=1}^{n} y_i \phi_i, \qquad i = 1, \ldots, n$$

we can write (5.13) in the following equivalent form

$$c(x^t B x)^{1/2} \leq \sup_{y \in \mathbb{R}^n} \frac{y^t A x}{(y^t B y)^{1/2}} \leq C(x^t B x)^{1/2}, \qquad \forall x \in \mathbb{R}^n.$$

Substituting $y = B^{-\frac{1}{2}} z$ inside the supremum above we get

$$\sup_{y \in \mathbb{R}^n} \frac{y^t A x}{(y^t B y)^{1/2}} = \sup_{z \in \mathbb{R}^n} \frac{z^t B^{-\frac{1}{2}} A x}{(z^t B^{-\frac{1}{2}} B B^{-\frac{1}{2}} z)^{1/2}} = |B^{-\frac{1}{2}} A x| = (x^t A^t B^{-\frac{1}{2}} B^{-\frac{1}{2}} A x)^{1/2},$$

where $|\cdot|$ denotes the Euclidean norm on $\mathbb{R}^n$. Thus, after squaring, we get the following equivalent form of (5.13)

$$c^2 x^t B x \leq x^t A^t B^{-1} A x \leq C^2 x^t B x, \qquad \forall x \in \mathbb{R}^n.$$

These estimates mean that $A^t B^{-1} A$ is spectrally equivalent to B. This suggests the following approach to solving the equation $Ax = b$: write the equation in the form

$$A^t B^{-1} A x = A^t B^{-1} b$$

and apply the preconditioned conjugate gradient (PCG) method to this system using B^{-1} as a preconditioner. Note that every PCG iteration will require the following number of matrix-vector multiplications: one with the matrix A, one with A^t, and two with the preconditioner B^{-1}. If we replace the matrix B with a spectrally equivalent matrix $\widetilde{B}$

$$c_1 x^t B x \leq x^t \widetilde{B} x \leq c_2 x^t B x, \qquad \forall x \in \mathbb{R}^n$$

then for all $x \in \mathbb{R}^n$ we have

$$\frac{c}{c_2}\,(x^t\widetilde{B}x)^{1/2} \le \frac{c}{\sqrt{c_2}}\,(x^tBx)^{1/2} \le \frac{1}{\sqrt{c_2}}\,\sup_{y\in\mathbb{R}^n}\frac{y^tAx}{(y^tBy)^{1/2}} \le \sup_{y\in\mathbb{R}^n}\frac{y^tAx}{(y^t\widetilde{B}y)^{1/2}} \quad \text{and}$$

$$\sup_{y\in\mathbb{R}^n}\frac{y^tAx}{(y^t\widetilde{B}y)^{1/2}} \le \frac{1}{\sqrt{c_1}}\,\sup_{y\in\mathbb{R}^n}\frac{y^tAx}{(y^tBy)^{1/2}} \le \frac{C}{\sqrt{c_1}}\,(x^tBx)^{1/2} \le \frac{C}{c_1}\,(x^t\widetilde{B}x)^{1/2}$$

and therefore the following spectral equivalence holds

$$(c/c_2)^2 x^t\widetilde{B}x \le x^tA^t\widetilde{B}^{-1}Ax \le (C/c_1)^2 x^t\widetilde{B}x, \qquad \forall x \in \mathbb{R}^n.$$

Note that the estimate for the condition number of $B^{-1}A^tB^{-1}A$ increased by a factor of $(c_2/c_1)^2$ when we replaced B with $\widetilde{B}$.

In the remainder of this section we describe how one can compute the action of B^{-1} or an appropriate $\widetilde{B}^{-1}$. Let $\{\phi\}_{i=1}^k$ be a basis for P_r on the reference element $\hat{T}$ such that (for example the standard nodal basis)

$$\sum_{i=1}^k \phi_i \equiv 1.$$

We introduce the new basis $\{\psi_i\}_{i=1}^k$

$$\psi_1 = \sum_{i=1}^k \phi_i \equiv 1, \qquad \psi_i = \phi_i - \alpha_i\psi_1, \quad i = 2,\ldots,k$$

where α_i are chosen so that $(\psi_i, 1)_{\hat{T}} = 0$; if we set $g_{ij} = (\phi_i, \phi_j)_{\hat{T}}$, $g_i = \sum_j g_{ij} = (\phi_i, 1)_{\hat{T}}$, and $g = \sum_i g_i = |\hat{T}|$, then $\alpha_i = g_i/g$. Note that $\{\psi_i\}_{i=1}^k$ form a basis since every ϕ_i can be expressed in terms of $\{\psi_j\}$

$$\phi_i = \psi_i + \alpha_i\psi_1, \quad i = 2,\ldots,k \qquad \phi_1 = \psi_1 - \sum_{i=2}^k \phi_i.$$

Define the $(k \times k)$ matrices $C = \{c_{ij}\}$ and $D = \{d_{ij}\}$ from the equalities

$$\psi_i = \sum_{j=1}^k c_{ij}\phi_j \qquad \phi_i = \sum_{j=1}^k d_{ij}\psi_j, \qquad i = 1,\ldots,k.$$

It is easy to see that $D\,C = I$ and that they have the form

$$C = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ -\alpha_2 & 1-\alpha_2 & & -\alpha_2 \\ \vdots & & \ddots & \\ -\alpha_k & -\alpha_k & & 1-\alpha_k \end{pmatrix} \qquad D = \begin{pmatrix} \alpha_1 & -1 & \cdots & -1 \\ \alpha_2 & 1 & & 0 \\ \vdots & & \ddots & \\ \alpha_k & 0 & & 1 \end{pmatrix}.$$

Let $T_1, \ldots, T_m$ be an enumeration of the elements in $\mathcal{T}$. We assume that the matrices A and B are defined based on the following basis for $\mathcal{V}$ (in the given order)

$$\Phi = \{\phi_1^1, \ldots, \phi_k^1, \phi_1^2, \ldots, \phi_k^2, \ldots, \phi_1^m, \ldots, \phi_k^m\}$$

where for $i = 1, \ldots, k$ and $j = 1, \ldots, m$

$$\phi_i^j(x) = \begin{cases} \phi_i(G_j^{-1}(x)), & x \in T_j \\ 0, & x \notin T_j \end{cases}$$

and G_j is the affine mapping from $\hat{T}$ to T_j. Let us also define the following ordered bases for $\mathcal{V}$

$$\Psi_1 = \{\psi_1^1, \ldots, \psi_k^1, \psi_1^2, \ldots, \psi_k^2, \ldots, \psi_1^m, \ldots, \psi_k^m\}$$
$$\Psi_2 = \{\psi_2^1, \ldots, \psi_k^1, \psi_2^2, \ldots, \psi_k^2, \ldots, \psi_2^m, \ldots, \psi_k^m, \psi_1^1, \ldots, \psi_1^m\}$$

where ψ_i^j is defined from ψ_i in the way ϕ_i^j was defined from ϕ_i. Note that for every element T_s we have

$$\psi_i^s = \sum_{j=1}^{k} c_{ij}\phi_j^s \qquad \phi_i^s = \sum_{j=1}^{k} d_{ij}\psi_j^s, \qquad i = 1, \ldots, k$$

$\psi_1^s|_{T_s} = 1$, and $(\psi_i^s, 1)_{T_s} = 0$, for $i = 2, \ldots, k$. Thus, the first $m(k-1)$ functions in Ψ_2 form a basis for $\mathcal{V}_0$ and the last m functions — a basis for $\overline{\mathcal{V}}$. If we denote by B_1 and B_2 the matrix representations of the inner product $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ in the bases Ψ_1 and Ψ_2,

respectively, then we have

$$B_1 = PB_2P^t,$$

where P is the permutation matrix that reorders the basis Ψ_2 into Ψ_1. Similarly, if $\widetilde{D}$ is the matrix of the coefficients from the representation of the basis Φ in terms of the basis Ψ_1, then

$$B = \widetilde{D}B_1\widetilde{D}^t.$$

Indeed, if $\Phi = \{\tilde{\phi}_i\}_{i=1}^n$, $\Psi_1 = \{\tilde{\psi}_i\}_{i=1}^n$, and $\tilde{\phi}_i = \sum_{j=1}^n \tilde{d}_{ij}\tilde{\psi}_j$ then

$$(B)_{ij} = \left\langle\!\left\langle \tilde{\phi}_j, \tilde{\phi}_i \right\rangle\!\right\rangle = \left\langle\!\left\langle \sum_{s=1}^n \tilde{d}_{js}\tilde{\psi}_s, \sum_{t=1}^n \tilde{d}_{it}\tilde{\psi}_t \right\rangle\!\right\rangle = \sum_{s,t=1}^n \tilde{d}_{js}\tilde{d}_{it}(B_1)_{ts} = (\widetilde{D}B_1\widetilde{D}^t)_{ij}.$$

In our case the matrix $\widetilde{D}$ is block-diagonal

$$\widetilde{D} = \begin{pmatrix} D & & 0 \\ & \ddots & \\ 0 & & D \end{pmatrix} \qquad \text{and} \qquad \widetilde{D}^{-1} = \widetilde{C} = \begin{pmatrix} C & & 0 \\ & \ddots & \\ 0 & & C \end{pmatrix}.$$

Using the above equalities for B and B_1 we get the following expression for B^{-1}

$$B^{-1} = (\widetilde{D}PB_2P^t\widetilde{D}^t)^{-1} = \widetilde{C}^tPB_2^{-1}P^t\widetilde{C}.$$

Since $\mathcal{V}_0$ and $\overline{\mathcal{V}}$ are orthogonal in the $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ inner product and

$$\langle\!\langle u_0, v_0 \rangle\!\rangle = (a\nabla u_0, \nabla v_0) \qquad \forall u_0, v_0 \in \mathcal{V}_0,$$

the matrix B_2 has the block-diagonal form

$$B_2 = \begin{pmatrix} B^{(1)} & & & 0 \\ & \ddots & & \\ & & B^{(m)} & \\ 0 & & & \overline{B} \end{pmatrix}$$

where $B^{(s)}$ is the $(k-1)\times(k-1)$ matrix with entries

$$(B^{(s)})_{ij} = \left\langle\!\left\langle \psi^s_{j+1}, \psi^s_{i+1} \right\rangle\!\right\rangle = (a\nabla\psi^s_{j+1}, \nabla\psi^s_{i+1}), \qquad i,j = 1,\dots,k-1$$

and $\overline{B}$ is the matrix representation of the bilinear form

$$\langle\!\langle \bar{u}, \bar{v} \rangle\!\rangle = \left\langle \kappa a_{\mathcal{E}} h_{\mathcal{E}}^{-1} [\![\bar{u}]\!], [\![\bar{v}]\!] \right\rangle_{\mathcal{E}_i \cup \mathcal{E}_D}, \qquad \bar{u}, \bar{v} \in \overline{\mathcal{V}}$$

in the standard (for $\overline{\mathcal{V}}$) basis $\{\psi^1_1, \dots, \psi^m_1\}$. The matrix $B^{(s)}$ can be computed using the equality

$$B^{(s)} = \hat{C}\hat{B}^{(s)}\hat{C}^t,$$

where $\hat{C}$ is the matrix obtained from C by removing the first row and

$$(\hat{B}^{(s)})_{ij} = (a\nabla\phi^s_j, \nabla\phi^s_i), \qquad i,j = 1,\dots,k.$$

Since every matrix-vector multiplication with B^{-1}

$$B^{-1} = \begin{pmatrix} C^t & & 0 \\ & \ddots & \\ 0 & & C^t \end{pmatrix} P \begin{pmatrix} (B^{(1)})^{-1} & & & 0 \\ & \ddots & & \\ & & (B^{(m)})^{-1} & \\ 0 & & & (\overline{B})^{-1} \end{pmatrix} P^t \begin{pmatrix} C & & 0 \\ & \ddots & \\ 0 & & C \end{pmatrix}$$

requires a matrix-vector multiplication with all matrices $(B^{(s)})^{-1}$ we can precompute and store them since they are of fixed small size. On the other hand $\overline{B}$ is an $(m\times m)$ matrix and its inverse is not sparse. Therefore we can assemble and store the sparse matrix $\overline{B}$ and then solve (e. g. using an iterative method) the equation $\overline{B}y = x$ every time we need to evaluate $y = (\overline{B})^{-1}x$. To avoid this exact (or almost exact) solve we can replace $(\overline{B})^{-1}$ with a preconditioner $\widehat{B}^{-1}$. Note that if $\widehat{B}$ is spectrally equivalent

to $\overline{B}$ then the resulting $\widetilde{B}$, defined by

$$\widetilde{B}^{-1} = \begin{pmatrix} C^t & & 0 \\ & \ddots & \\ 0 & & C^t \end{pmatrix} P \begin{pmatrix} (B^{(1)})^{-1} & & & 0 \\ & \ddots & & \\ & & (B^{(m)})^{-1} & \\ 0 & & & \widehat{B}^{-1} \end{pmatrix} P^t \begin{pmatrix} C & & 0 \\ & \ddots & \\ 0 & & C \end{pmatrix},$$

will be spectrally equivalent to B. To define a preconditioner $\widehat{B}$ one can use a multigrid algorithm based on the sequence of nested spaces (using the notation introduced on page 39)

$$\overline{\mathcal{V}}_1 \subset \overline{\mathcal{V}}_2 \subset \cdots \subset \overline{\mathcal{V}}_J = \overline{\mathcal{V}}$$

and the bilinear forms

$$\langle\!\langle \bar{u}, \bar{v} \rangle\!\rangle_k = \left\langle \kappa h_{\mathcal{E}_k}^{-1} a_{\mathcal{E}_k} [\![\bar{u}]\!], [\![\bar{v}]\!] \right\rangle_{\mathcal{E}_i^k \cup \mathcal{E}_D^k}, \quad \forall \bar{u}, \bar{v} \in \overline{\mathcal{V}}_k.$$

Note that this method is almost the same as the multigrid Method II used for the preconditioing of the SIPG method (see page 41). The only difference is that Method II has one additional level $M_{J+1} = \mathcal{V}_J = \mathcal{V}$. On all other levels the spaces are the same and the bilinear forms coincide since $\langle\!\langle \cdot, \cdot \rangle\!\rangle_k$ is equal to the k-th level SIPG bilinear form on the space $\overline{\mathcal{V}}_k$.

5.5. Numerical Experiments

In this section we present numerical experiments using the preconditioners described in the previous section. The majority of the tests use quadratic finite elements since this is the case for which we proved that the inf-sup condition (5.3) holds independently of h and as a consequence B and $\widetilde{B}$ are uniform preconditioners in this case. We use W-cycle multigrid algorithm (as described above) with symmetric Gauss-Seidel

smoothing to define the preconditioner $\widehat{B}$ used in the definition of $\widetilde{B}$. We tested the following combinations of iterative solvers and preconditioners:

- M1: apply the PCG method to the symmetrized system $A^t B^{-1} A x = A^t B^{-1} b$ using B^{-1} as the preconditioner. This is a "two-level"-type method since we need to solve a coarse problem to define the action of B^{-1}.

- M2: apply the PCG method to the symmetrized system $A^t \widetilde{B}^{-1} A x = A^t \widetilde{B}^{-1} b$ using $\widetilde{B}^{-1}$ as the preconditioner.

- M3: apply the GMRES method to the original system $Ax = b$ using $\widetilde{B}^{-1}$ as the preconditioner.

- M4: apply the GMRES(10) method (GMRES restarted every 10 iterations) to the original system $Ax = b$ using $\widetilde{B}^{-1}$ as the preconditioner.

The linear systems are solved with the same relative accuracy of 10^{-8}, that is the stopping criterion is $(e_i/e_0) < 10^{-8}$ where $e_i^2 = r_i^t M r_i$ for PCG and $e_i^2 = r_i^t M^t M r_i$ for GMRES where r_i is the i-th residual and M is the preconditioner. We discretize and solve the same Test Problems as in the previous chapters which are described on page 32. As we did earlier, for each domain we generate a sequence of nested meshes starting with a coarse ("Level 0") tetrahedral mesh and then using k times uniform refinement to obtain the "Level k" mesh.

In Table 5.2 we present the results for Test Problem 1 using quadratic elements. For these tests the parameter κ in the definition of $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ was chosen to be $\kappa = 1$. Note that, in contrast to the SIPG method, the choice of κ affects the preconditioner but not the matrix A. The first two rows in the table give the number of degrees of freedom (dof) in the discretization space $\mathcal{V}$ and the piecewise constant space $\overline{\mathcal{V}}$. The other rows give the number of iterations for the corresponding solution methods M1–

Table 5.2. Preconditioners for the method of Baumann and Oden, Test Problem 1, quadratic FE

	Level 1	Level 2	Level 3	Level 4	Level 5
$\mathcal{V}$ dof	960	7,680	61,440	491,520	3,932,160
$\overline{\mathcal{V}}$ dof	96	768	6,144	49,152	393,216
M1	22	46	53	53	52
M2	37	55	78	86	88
M3	54	79	88	87	85
M4	87	109	108	107	103

M4. From these results we see that in all cases the iteration counts remain bounded with the refinement of the mesh. Comparing the results for M1 and M2 we see that replacing the exact coarse solve in B (method M1) with the multigrid preconditioner in $\widetilde{B}$ (method M2) results in about 70% increase in the number of iterations in the worst case. If we compare the results for M2 and M3 we see that they both converge for about the same number of iterations. However, each iteration in the PCG method (M2) requires two matrix-vector multiplications with $\widetilde{B}^{-1}$, one with A and one with A^t, whereas the GMRES method (M3) requires just one action of $\widetilde{B}^{-1}$ and one of A. On the other hand every iteration in the GMRES method requires an increasing number of vector updates — every new iteration uses one additional vector update compared to the previous iteration. One way of dealing with this problem is to restart the GMRES method after a certain number of iterations. With this approach the number of vector updates remains small but usually the convergence is slower. In our tests using GMRES(10) (method M4) we see an increase in the number of iterations of about 20–25% compared to the standard GMRES (method M3).

Table 5.3. Preconditioners for the method of Baumann and Oden, Test Problem 2, quadratic FE

	Level 1	Level 2	Level 3	Level 4
$\mathcal{V}$ dof	3,360	26,880	215,040	1,720,320
$\overline{\mathcal{V}}$ dof	336	2,688	21,504	172,032
M1, $\epsilon = 1$	37	40	40	40
M1, $\epsilon = 0.1$	56	69	70	70
M1, $\epsilon = 0.01$	241	401	462	461
M2, $\epsilon = 1$	40	56	65	66
M3, $\epsilon = 1$	64	68	68	65
M4, $\epsilon = 1$	76	77	78	76

In Table 5.3 we present the results for Test Problem 2 using quadratic elements. Here, as in the previous Test Problem, we took $\kappa = 1$. The first two rows in the table give the number of degrees of freedom (dof) for the spaces $\mathcal{V}$ and $\overline{\mathcal{V}}$. The rest of the rows give the number of iterations for the indicated solution method and value of the parameter ϵ. In all cases we observe that the iterations remain bounded as we refine the mesh. However, decreasing the value of ϵ results in a substantial increase in the number of iterations as seen from the results for method M1. Since we observe this effect when using the preconditioner B, we can not expect better results for $\widetilde{B}$ (methods M2–M4). This numerical results confirm, as indicated in Remark 9, that the estimate of Lemma 7 depends on the jumps of the coefficient a across interior faces (note that the other Lemmas 8 and 10 leading to the inf-sup condition and boundedness estimates (5.13) are independent of such jumps). The rest of the results in Table 5.3 are for the case $\epsilon = 1$ and we observe a behavior very similar to Test Problem 1 with slightly better convergence rates for all methods.

Table 5.4. Preconditioners for the method of Baumann and Oden, varying κ, Test Problem 1, quadratic FE

	Level 1	Level 2	Level 3	Level 4	Level 5
M4, $\kappa = 0.25$	149	169	168	180	166
M4, $\kappa = 0.5$	108	129	127	130	124
M4, $\kappa = 1$	87	109	108	107	103
M4, $\kappa = 2$	73	103	103	103	102
M4, $\kappa = 4$	68	107	110	110	107
M4, $\kappa = 8$	87	144	128	137	131

With the next set of numerical experiments we test the effect of varying the parameter κ. The results for Test Problem 1 using method M4 are presented in Table 5.4. For this case we see that the optimal choice is around $\kappa = 2$. Increasing or decreasing the value of κ away from this optimal value results in an increase in the number of iterations. The results show that the optimal value is fairly independent of the refinement level. Therefore one can use coarse problems to determine a value of κ close the optimal that can be used for the large problems.

In the final set of numerical experiments presented here we use linear finite elements to discretize Test Problem 1. All the meshes we used (levels 2–6) are checkerboard and since $\Gamma_D = \partial\Omega$ the discrete linear systems are singular and the kernel of the operators A and A^t consists of all checkerboard functions (see Lemma 12 and Remark 10). With the right hand side of Test Problem 1 ($f = 1$) and the triangulations we use we have that $\mathcal{L}(v) = 0$, $\forall v \in \operatorname{Ker} A^t$, i. e. the discrete linear equations are compatible. We apply method M1 with the zero vector as initial approximation of the solution. Even though the matrices of the linear systems are singular the PCG method converges in all cases we tested. The number of iterations for various values

Table 5.5. Preconditioners for the method of Baumann and Oden, Test Problem 1, linear FE

	Level 2	Level 3	Level 4	Level 5	Level 6
M1, $\kappa = 0.025$	52	79	83	79	74
M1, $\kappa = 0.05$	49	70	68	66	65
M1, $\kappa = 0.1$	48	63	63	64	58
M1, $\kappa = 0.2$	46	64	66	64	62
M1, $\kappa = 0.4$	44	72	77	74	69
M1, $\kappa = 0.8$	46	90	96	90	83

of κ are listed in Table 5.5. In all these tests the solution we obtain is orthogonal (as a vector in $\mathbb{R}^n$) to the vector representation of the checkerboard function. This means that the solution is L^2-orthogonal to the checkerboard functions since the tetrahedra in each of our meshes have the same volume. Unfortunately, we do not have a rigorous explanation of this observation. Our theoretical results do not cover the case of linear elements. However, the numerical results clearly show that the number of iterations remains bounded as we refine the mesh. Also, we see that the optimal value for κ is around 0.1 which is much smaller than the optimal value for method M4 with quadratic elements (our previous test). Even though it is natural to expect that replacing the preconditioner B in method M1 with $\widetilde{B}$ (method M2) will still give uniform (with respect to h) convergence, this is not confirmed by our tests. In fact, when using method M2, we observe a substantial increase in the number of iterations with the refinement of the mesh.

CHAPTER VI

SUMMARY

We have developed and numerically tested a number of preconditioners for two discontinuous Galerkin (DG) methods for second order elliptic problems. The two DG methods are the symmetric interior penalty (SIPG) method and the method of Baumann and Oden.

In Chapter III we introduced two- and multilevel preconditioners for the SIPG method based on two types of coarse spaces consisting of (1) continuous piecewise polynomial or (2) piecewise constant functions. We proved that both two-level preconditioners and the multilevel preconditioner based on continuous coarse spaces give convergence rates independent of the mesh size. The presented numerical experiments confirm these results. Even though we do not have theoretical results for the multilevel preconditioner based on piecewise constant coarse spaces, in the numerical experiments we observe uniform convergence for the W-cycle.

In Chapter IV we introduced an algebraic multigrid (AMG) preconditioner for the SIPG method that uses coarsening based on element agglomeration. We also considered a version of the algorithm using a smoothed aggregation technique designed to improve the convergence. A major advantage of these AMG methods over the methods of Chapter III is the fact that they can be used on unstructured meshes. We do not have theoretical analysis for the proposed AMG preconditioners, however the numerical experiments we presented showed that they give uniform or almost uniform convergence rates.

In Chapter V we presented an approach for constructing a preconditioner for the method of Baumann and Oden. We proved that this preconditioner is spectrally equivalent to an appropriate symmetrization of the discrete linear system when the

finite elements used are quadratic or higher order. In the case of linear finite elements we gave a characterization of the kernel of the discrete operator and presented numerical experiments showing optimal convergence rates for the DG method in both L^2 and H^1 norms. The numerical results presented in the end of the chapter confirmed the theoretical result that the proposed preconditioner gives convergence rates independent of the element size when used in a PCG iteration applied to the symmetrized discrete linear system. In addition, we observed similar behavior when the preconditioner was used in a GMRES or restarted GMRES iteration applied to the original linear system.

REFERENCES

[1] D. Arnold, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.

[2] D. Arnold, F. Brezzi, B. Cockburn, and D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[3] I. Babuška, and M. Zlámal, *Nonconforming elements in the finite element method with penalty*, SIAM J. Numer. Anal., 10 (1973), pp. 863–875.

[4] F. Bassi and S. Rebay, *A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations*, J. Comput. Phys., 131 (1997), pp. 267–279.

[5] C. E. Baumann and J. T. Oden, *A discontinuous hp finite element method for convection-diffusion problems*, Comput. Methods Appl. Mech. Engrg., 175 (1999), pp. 311–341.

[6] J. H. Bramble, J. E. Pasciak, and and J. Xu, *The analysis of multigrid algorithms with nonnested spaces or noninherited quadratic forms*, Math. Comp., 56 (1991), pp. 1–34.

[7] J. H. Bramble and J. E. Pasciak, *The analysis of smoothers for multigrid algorithms*, Math. Comp., 58 (1992), pp. 467–488.

[8] J. H. Bramble and X. Zhang, *The analysis of multigrid methods*, in Handbook of Numerical Analysis, P. G. Ciarlet and J. L. Lions, eds., Elsevier Science, Amsterdam, 2000, vol. VII, pp. 173–416.

[9] S. Brenner and J. Zhao, *Convergence of multigrid algorithms for interior penalty methods*, Appl. Numer. Anal. Comput. Math., 2 (2005), pp. 3–18.

[10] M. Brezina, A. J. Cleary, R. D. Falgout, V. E. Henson, J. E. Jones, T. A. Manteuffel, S. F. McCormick, and J. W. Ruge, *Algebraic multigrid based on element interpolation (AMGe)*, SIAM J. Sci. Comput., 22 (2000), pp. 1570–1592.

[11] F. Brezzi and M. Fortin, *Mixed and hybrid finite element methods*, Springer Series in Computational Mathematics, 15. Springer-Verlag, New York, 1991.

[12] F. Brezzi, G. Manzini, D. Marini, P. Pietra, and A. Russo, *Discontinuous finite elements for diffusion problems*, in Atti Convegno in onore di F. Brioschi (Milan, 1997), Instituto Lombardo, Accademia di Scienze e Lettere, Milan, Italy, 1999, pp. 197–217.

[13] K. Brix, M. Campos Pinto, W. Dahmen, *A multilevel preconditioner for the interior penalty discontinuous Galerkin method*, IGPM Report, RWTH Aachen, 2007, http://www.igpm.rwth-aachen.de/dahmen/BCD_final.pdf

[14] P. G. Ciarlet, *The finite element method for elliptic problems*, North-Holland, Amsterdam, 1978.

[15] B. Cockburn and C.-W. Shu, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.

[16] V. Dobrev, R. Lazarov, P. Vassilevski, and L. Zikatanov, *Two-level preconditioning of discontinuous Galerkin approximations of second-order elliptic equations*, Numer. Linear Algebra Appl., 13 (2006), pp. 753–770.

[17] V. A. Dobrev, R. D. Lazarov, and L. T. Zikatanov, *Preconditioning of symmetric interior penalty discontinuous Galerkin FEM for second order elliptic problems*, in Proc. Int. Conf. on Domain Decomposition Methods, DD-17, Strobl, Austria, (2007), to appear.

[18] J. Douglas, Jr. and T. Dupont *Interior penalty procedures for elliptic and parabolic Galerkin methods*, Lecture Notes in Phys. 58, Springer-Verlag, Berlin, 1976.

[19] J. Gopalakrishnan, and G. Kanschat, *A multilevel discontinuous Galerkin method*, Numer. Math., 95 (2003), pp. 527–550.

[20] J. Jones and P. Vassilevski *AMGe based on element agglomeration*, SIAM J. Sci. Comput., 23 (2001), pp. 109–133.

[21] J. T. Oden, I. Babuška, and C. E. Baumann, *A discontinuous hp finite element method for diffusion problems*, J. Comput. Phys., 146 (1998), pp. 491–519.

[22] B. Rivière, M. Wheeler, and V. Girault, *Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems I*, Comput. Geosci., 3 (1999), pp. 337–360.

[23] T. Rusten, P. Vassilevski, and R. Winther, *Interior penalty preconditioners for mixed finite element approximations of elliptic problems*, Math. Comp., 65 (1996), pp. 447–466.

[24] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.

[25] P. Vaněk, *Acceleration of convergence of a two-level algorithm by smoothing transfer operators*, Appl. Math., 37 (1992), pp. 265–274.

[26] P. Vaněk, M. Brezina, and J. Mandel, *Convergence of algebraic multigrid based on smoothed aggregation*, Numer. Math., 88 (2001), pp. 559–579.

[27] M. F. Wheeler, *An elliptic collocation-finite element method with interior penalties*, SIAM J. Numer. Anal., 15 (1978), pp. 152–161.

VITA

Veselin Asenov Dobrev was born in Sofia, Bulgaria. He studied at Sofia University "St. Kliment Ohridski" from September 1993 and graduated in September 1998 with a Bachelor of Science degree in Applied Mathematics. After his graduation Veselin worked as a researcher in the Central Laboratory for Parallel Processing (now Institute for Parallel Processing) in the Bulgarian Academy of Sciences. In the fall of 2000 he enrolled in the doctoral program in the Department of Mathematics, Texas A&M University. The current dissertation on preconditioning of discontinuous Galerkin methods was defended in May 2007. Veselin can be contacted at:

Department of Mathematics

Texas A&M University

College Station, TX 77843-3368

U. S. A.

E-mail address: dobrev@math.tamu.edu

The typist for this dissertation was Veselin Dobrev.